COPYRIGHT AND USE OF THIS THESIS

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author’s moral rights if you:
- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author’s reputation

For further information contact the University’s Director of Copyright Services

sydney.edu.au/copyright
The interdependence between biological and ethical analyses of safety and efficacy in translational gene therapy

Claire Tara Deakin

Submitted in fulfilment of the requirements for admission to the degree of Doctor of Philosophy

The Children's Medical Research Institute

Sydney Medical School

The University of Sydney

April 2013
DECLARATION

The work described in this thesis was performed at the Children’s Medical Research Institute, under the supervision of Professor Ian Alexander, Dr Samantha Ginn, Dr Claus Hallwirth and Associate Professor Ian Kerridge. All work was performed personally by the author, Claire Deakin, unless acknowledged otherwise. This work has not been previously submitted for any degree at any institution.

Claire Tara Deakin

18 April 2013
This research project stems from the successful clinical trials of gene therapy for X-linked severe combined immunodeficiency (SCID-X1). Whilst those trials represented the first successful clinical application of gene therapy for a genetic disease, the subsequent development of leukaemia in 5 of 20 patients pointed to the need to develop safer vectors for gene transfer that retain a high degree of efficacy. Furthermore, the unanticipated frequency of these serious adverse events highlights the uncertainty of risk in early phase human trials and the challenges of assessing uncertain risks when trials are designed. To address these technological and ethical challenges, this project has taken a unique, interdisciplinary approach, adopting both a molecular, laboratory-based component and an ethical, social sciences-based component.

The laboratory component of this project focussed on the development of lentiviral vectors for haematopoietic gene therapy, such as gene therapy for SCID-X1. Specifically, the impact of features of vector design, such as the selection of an appropriate promoter-enhancer element for controlling expression of a therapeutic transgene, on the clonal complexity of reconstituted lymphocytes and on the risk of lymphomagenesis due to proliferative stress were investigated using a murine model of SCID-X1. Although the selection of a less transcriptionally active promoter-enhancer was associated with reduced lymphocyte reconstitution, ultimately there was no difference in the incidence of lymphomagenesis in mice treated with either of 2 vectors containing different promoter-enhancer elements. Analysis of the clonal complexity of lymphocytes in mice treated with either of the 2 vectors, based on identification of vector integration sites, was inconclusive. Therefore, the hypothesis that use of a less
transcriptionally active promoter-enhancer element is associated with reduced clonal complexity cannot be definitively excluded.

Additionally, the laboratory component of this project explored the feasibility of vector barcoding coupled with next generation sequencing analysis as a potential tool for measuring and monitoring clonal complexity and dynamics in preclinical and clinical studies. While complex barcoded plasmid libraries were successfully generated, sequencing error prevented analysis of the exact complexity of those libraries. Analyses of samples containing 1, 10 and 100 known barcode sequences indicated that true barcode sequences could be distinguished from background when complexity was low, at 1 or 10 barcode combinations, but not when complexity was moderate, at 100 barcode combinations. Therefore, the degree of library complexity that can be resolved using next generation sequencing was insufficient to be clinically useful. Nonetheless, insights were uncovered with regard to the nature of systematic sequencing error that gave rise to high frequency false barcode sequences.

The ethics component of this project has given attention to the meaning of risk and how knowledge about risk is constructed, perceived and assessed in the context of translational gene therapy. The manner in which gene therapy researchers perceive and assess risks was investigated empirically. Researchers’ decision-making depended strongly on both the strength of the available preclinical evidence and the nature of the clinical context for the disease under consideration. A number of subjective influences on decision-making were found, including attitudes toward the strength and validity of preclinical data, perceptions of the need for a novel treatment, personal experience of disease, length and kind of professional experience and attitudes toward serious adverse events.
By evaluating the impact of improved vector design on the safety and efficacy of SCID-X1 gene therapy, working towards the development of tools with greater sensitivity for measuring efficacy and monitoring safety, and investigating how gene therapy researchers assess and perceive risks, this project represents a novel and unique contribution to how safety and efficacy are understood in translational gene therapy.
ACKNOWLEDGEMENTS

I have been privileged to meet, work with and learn from some remarkable individuals during the past 5 years. In particular, I am grateful for the consistent support and encouragement I have received from my supervisors, Professor Ian Alexander, Dr Samantha Ginn, Dr Claus Hallwirth and Associate Professor Ian Kerridge, and I thank them for their feedback on earlier drafts of this thesis. I would like to thank Ian Alexander for supporting my passion for taking on an interdisciplinary project, training me to be a better scientist, encouraging me to pursue my own ideas, teaching me to be a more effective written and oral communicator, and giving me so much of his time. I thank Sam for all the laboratory skills she taught me, particularly everything she taught me about lentiviral vectors for X-linked severe combined immunodeficiency. I thank Claus for sharing his insights into molecular biology and next generation sequencing, and for making time to discuss experiment designs.

I would also like to thank Ian Kerridge for mentoring me as an interdisciplinary researcher, for guiding the ethics project, and for making so much time for me. I am also grateful to his colleagues at the Centre for Values, Ethics and the Law in Medicine, Professor Cliff Hooker and Dr Claire Hooker, for their insights and critical feedback.

All of my colleagues in the Gene Therapy Research Unit have been wonderful to work with during the past 5 years and I thank them all for the technical and moral support over the years. In particular, I would like to thank Sophia Liao for her assistance with the mouse work and Dr Christine Smyth for advice about flow cytometry. I am also grateful to Dr Belinda Kramer of the Children’s Cancer Research Unit, Kids’ Research Institute, for the advice she has given me.
A number of researchers and staff at the Children’s Medical Research Institute have helped me over the years, and I have relished working in such a friendly, supportive and collaborative environment. I thank Shelley Dimech for the care she has shown in looking after the immunocompromised mice studied in this thesis, and particularly for helping to diagnose sick mice. I thank Dr Afra Engelman for passing on her method for ligation-mediated PCR and Dr Hamish Campbell for his help with reading an earlier draft of this thesis. I thank Associate Professor Catherine Suter, Dr David Humphries and Paul Young from the Victor Chang Cardiac Research Institute for their assistance with sequencing on the SOLiD platform and data analysis.

I am grateful for all the help I have received with bioinformatic analysis. Most particularly, I thank my father for lending his years of programming experience to the barcoding project and for taking an active and engaged role in data analysis. I would also like to thank Dr Gagan Garg from the Gene Therapy Research Unit and Mohammad Islam from Macquarie University for their help with mapping vector integration sites. I also thank Adam Skarszewski from the Australian Genome Research Facility for his help with the barcoding project at the earlier stages.

For financial support, I am grateful to have received a Yass Memorial Scholarship from the Children’s Medical Research Institute and an National Health and Medical Research Council Biomedical Postgraduate Scholarship. I thank the Wenkart Foundation for funding my reagents and consumables during 2010–2011.

Finally, I would like to thank my friends for their moral support and for the moments of diversion from laboratory experiments, and I thank my family, particularly my parents, for the love, support and encouragement they give me in everything I do.
PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS WORK

Manuscripts

(Published manuscripts are reproduced in Appendix 3)

(1) Deakin CT, Deakin JJ, Hallwirth CV, Ginn SL, Alexander IE, “Impact of next generation sequencing error upon the analysis of barcoded plasmid libraries of defined complexity comprising known sequence identities” (Manuscript in preparation)

(2) Deakin CT, Alexander IE, Hooker CA and Kerridge IH, “Gene therapy researchers’ assessments of risks and perceptions of risk acceptability in clinical trials”, Molecular Therapy (2013) 21(4): 806-15

(Publication of the results described in Chapter 5)


(Invited review, on the basis of the review published in Molecular Therapy in 2009)


Oral Presentations:

(1) Deakin CT, Deakin JJ, Hallwirth CV, Ginn SL, Young P, Humphries D, Suter CM and Alexander IE, “Impact of next generation sequencing error on the analysis of barcoded plasmid libraries of defined complexity comprising known sequence identities”, 8th Australasian Gene Therapy Society Meeting in Sydney, Australia, 8-10 May 2013

(2) Deakin CT, “Investigating the feasibility of next generation sequencing for analysing the complexity of molecular barcode libraries”, Sydney Next Generation Sequencing Special Interest Group Meeting in Sydney, Australia, 29 November 2012
Poster Presentations:


(3) Deakin CT, Ginn SL, Hallwirth CV, Skarszewski A and Alexander IE, “Development of a lentiviral barcoding methodology and analysis of a complex barcoded plasmid library using next generation sequencing”, 16th Human Genome Meeting in Sydney, Australia, 11-14 March 2012

(4) Deakin CT, Ginn SL, Liao SH and Alexander IE), “Construction and validation of a barcoded lentivirus stock for investigating safety and efficacy of gene therapy for X-linked severe combined immunodeficiency”, 7th Australasian Gene Therapy Society Meeting in Melbourne, Australia, 4-6 May 2011; also presented at Westmead Hospital Week in Sydney, Australia, 12-16 September 2011

(Awarded Australasian Gene Therapy Society Student Prize)

(5) Deakin CT, Alexander IE and Kerridge IH, “Investigating gene therapy researchers’ assessment of risks in clinical trials”, 7th Australasian Gene Therapy Society Meeting in Melbourne, Australia, 4-6 May 2011

(Awarded Australasian Gene Therapy Society Student Prize)


(Awarded Australasian Gene Therapy Society Student Prize)
CHAPTER ONE: INTRODUCTION

1.1 SAFETY AND EFFICACY IN TRANSLATIONAL GENE THERAPY

1.2 DEVELOPMENT OF GENE THERAPY FOR X-LINKED SEVERE COMBINED IMMUNODEFICIENCY (SCID-X1)

1.2.1 Background on SCID-X1

1.2.2 Limitations of existing treatment for SCID-X1 and rationale for development of gene therapy

1.2.3 Lessons for SCID-X1 gene therapy from spontaneous reversion events and partial SCID-X1 phenotypes

1.2.4 Strategy for SCID-X1 gene therapy

1.2.5 Trials of haematopoietic gene therapy targeting other diseases

1.3 CLINICAL TRIALS OF GENE THERAPY FOR SCID-X1

1.3.1 Efficacy of SCID-X1 gene therapy in the French and British clinical trials

1.3.2 Occurrence of serious adverse events in five patients across the two independent SCID-X1 trials
1.3.3 Vector-mediated insertional mutagenesis as the causative mechanism of leukaemogenesis 20

1.3.4 Role of LMO2 in lymphomagenesis 22

1.3.5 Possible impact of SCID-X1 background and biology on insertional mutagenesis risks 22

1.3.6 Insertional mutagenesis in other gene therapy trials 24

1.4 LIMITATIONS OF PRECLINICAL MODELS FOR ASSESSING GENOTOXICITY 25

1.4.1 Assessments of insertional mutagenesis prior to serious adverse events in the SCID-X1 trials 26

1.4.2 Limitations of preclinical models 28

1.4.3 Development of improved preclinical models 30

1.4.4 Development of knowledge about insertional mutagenesis risks and the interpretation of read-outs from preclinical models 32

1.5 ANALYSIS OF THE INTEGRATION BEHAVIOUR OF RETROVIRAL VECTORS 33

1.5.1 Integration of gammaretroviral vectors is non-random 34

1.5.2 Gammaretroviral integration can trigger clonal dominance 42

1.5.3 Gammaretroviral vectors preferentially integrate near transcriptional start sites compared to lentiviral vectors 43

1.5.4 Limitations of methods for analysing retroviral integrations 45

1.5.5 Development of non-restrictive methods 47

1.5.6 Application of next generation sequencing technology to integration site analysis 48

1.6 STRATEGIES FOR DESIGNING SAFER VECTORS FOR SCID-X1 GENE THERAPY 49

1.6.1 Self-inactivating vector configuration 49

1.6.2 Development of HIV-1-based lentiviral vectors for gene therapy 51

1.6.3 Modification of the HIV-1 genome to produce replication incompetent lentiviral vectors for gene therapy 52

1.6.4 Existing data evaluating the safety of lentiviral vectors compared to gammaretroviral vectors 56
2.3.3 Routine nucleic acid procedures

2.3.3.1 Estimation of nucleic acid concentration and purity

2.3.3.2 Agarose gel electrophoresis of DNA

2.3.3.3 Polyacrylamide gel electrophoresis of DNA

2.3.3.4 Column-based DNA purification

2.3.3.5 Phenol:chloroform extraction of DNA

2.3.3.6 Ethanol precipitation of DNA

2.3.4 Transformation of competent bacteria with plasmid DNA

2.3.4.1 Bacterial transformation using the heat-shock method

2.3.4.2 Bacterial transformation using electroporation

2.3.5 Recombinant DNA techniques

2.3.5.1 Restriction endonuclease digestion of DNA

2.3.5.2 Ligation of DNA molecules

2.3.6 Subcloning of PCR products

2.3.7 Sanger sequencing of DNA

CHAPTER THREE: EVALUATING THE SAFETY AND EFFICACY OF LENTIVIRAL VECTORS WITH γc EXPRESSION UNDER THE TRANSCRIPTIONAL CONTROL OF DIFFERENT PROMOTER-ENHANCER ELEMENTS

3.1 INTRODUCTION

3.2 CHAPTER 3 HYPOTHESIS AND AIMS

3.3 MATERIALS AND METHODS

3.3.1 Lentiviral vector production

3.3.1.1 Plasmid constructs used in lentiviral vector production

3.3.1.2 Transient co-transfection of lentiviral packaging and transfer vector plasmids

3.3.1.3 Lentiviral vector harvest and concentration

3.3.1.4 Assignment of transducing titre to lentiviral stock
3.3.2 Treatment of γc-deficient mice with genetically-modified Sca1⁺ bone marrow cells

3.3.2.1 Mouse strains used for the murine model of SCID-X1 gene therapy

3.3.2.2 Isolation and transduction of IL2Rγ⁻/⁻ Sca1⁺ bone marrow cells for transplantation into IL2Rγ⁻/⁻ Rag2⁻/⁻ recipient mice

3.3.2.3 Evaluation of immune reconstitution by flow cytometric analysis of lymphocyte markers

3.3.3 Macroscopic examination and analysis of cellular characteristics of murine lymphomas

3.3.4 Analysis of molecular characteristics of murine lymphomas

3.3.4.1 qPCR analysis of vector copy number

3.3.4.2 Integration site analysis by conventional ligation-mediated PCR (LM-PCR)

3.3.4.3 Microarray analysis of gene expression

3.3.5 Preparation of vector-genomic junction fragment libraries representing lentiviral vector integration sites for next generation sequencing

3.3.5.1 Extraction and digestion of genomic DNA from EF1α.γc-treated and PGK.γc-treated mice

3.3.5.2 Preparation and ligation of a linker cassette to Tsp509I-digested genomic DNA

3.3.5.3 Initial PCR amplification of vector-genomic DNA

3.3.5.4 Nested PCR amplification of vector-genomic DNA

3.3.6 Next generation sequencing of vector-genomic DNA libraries and data analysis

3.3.7 Statistical analysis

3.4 RESULTS

3.4.1 Setting up of cohorts of IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

3.4.2 Reconstitution of immunity in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors
3.4.3 Skewed T cell to B cell ratio and CD4⁺ T cell to CD8⁺ T cell ratio in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

3.4.4 Long-term safety monitoring of IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

3.4.4.1 Macroscopic and cellular characteristics of lymphomas arising in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

3.4.4.2 Analyses of vector copy numbers, vector integration sites and gene dysregulation in lymphomas

3.4.5 Selection of representative samples for preparation of libraries vector-genomic DNA fragment libraries

3.4.6 Sequencing of integration site libraries and data processing

3.4.7 Mapping of unique integration sites in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

3.5 DISCUSSION OF CHAPTER 3 RESULTS

3.5.1 No significant difference in rates of lymphomagenesis between the EF1α.γc and PGK.γc lentiviral vectors EF1α.γc and PGK.γc lentiviral vectors

3.5.2 Inconclusive results of the integration site analysis for comparing clonal complexity in mice treated with the EF1α.γc and PGK.γc lentiviral vectors

3.5.3 Impaired reconstitution of NK cell and B cell subsets following treatment with the EF1α.γc and PGK.γc lentiviral vectors

3.5.4 Limitations of the molecular analyses of lymphomas

3.6 CONCLUSIONS FOR CHAPTER 3

CHAPTER FOUR: GENERATION OF LIBRARIES OF BARCODED VECTORS FOR ANALYSING CLONAL COMPLEXITY AND EVALUATION OF LIBRARY COMPLEXITY USING NEXT GENERATION SEQUENCING

4.1 INTRODUCTION

4.2 CHAPTER 4 HYPOTHESIS AND AIMS

4.3 MATERIALS AND METHODS

4.3.1 Generation of complex barcoded plasmid libraries
4.3.1.1 Primer extension methodology to construct double-stranded barcode inserts

4.3.1.2 Cloning of double-stranded barcode inserts into pEF1α.γc for generating a complex barcoded plasmid library

4.3.1.3 Small-scale validation of library complexity

4.3.2 Next generation sequencing of barcode libraries

4.3.3 Data analysis

4.3.3.1 Data analysis for the initial barcode design

4.3.3.2 Data analysis for the Illumina-specific barcode design

4.3.3.3 Data analysis for the SOLiD-specific barcode design

4.3.3.4 Analyses of empirical error rates and error types

4.3.3.5 Analyses of sequence characteristics

4.3.4 Empirical assessment of individual barcode representation within the Illumina-specific 100-barcode library

4.3.5 Statistical analysis

4.4 RESULTS

4.4.1 Construction and validation of an initial barcoded pEF1α.γc lentiviral plasmid library

4.4.2 Analysis of the initial barcoded pEF1α.γc lentiviral plasmid library using Illumina sequencing

4.4.2.1 Preparation of barcode amplicons for next generation sequencing

4.4.2.2 Analysis of a sample containing 2 known barcodes

4.4.2.3 Development of bioinformatic strategies for reducing background

4.4.2.4 Analysis of the complex libraries

4.4.2.5 Analyses of empirical error frequencies and quality of sequence reads

4.4.3 Design, construction and preliminary validation of platform-specific barcoded pEF1α.γc lentiviral plasmid libraries

4.4.4 Analyses of barcode libraries of defined complexity containing 1, 10
or 100 known barcodes

4.4.4.1 Preparation of barcode libraries of defined complexity 177

4.4.4.2 Sequencing of Illumina-specific barcode libraries of defined complexity 178

4.4.4.3 Analyses of the Illumina-specific 1-barcode sample 178

4.4.4.4 Analysis of the Illumina-specific 10-barcode library 181

4.4.4.5 Analyses of the Illumina-specific 100-barcode library using 2 independent sequencing runs 181

4.4.4.6 Analysis of putatively false barcodes detected in the Illumina-specific 100-barcode library 184

4.4.4.7 Analysis of the SOLiD-specific 100-barcode library 184

4.4.5 Validation of the Illumina-specific 100-barcode library 187

4.4.5.1 Discrepancy between the actual and detected abundance of barcodes 187

4.4.6 Characterisation of empirical error rate, error location and substitution error type for sequencing runs using the Illumina platform 192

4.4.7 Analyses of a highly complex barcode library of unknown complexity 195

4.4.7.1 Sequencing of the highly complex Illumina-specific barcode library 195

4.4.7.2 Serial dilution approach to analysing the complexity of the highly complex Illumina-specific barcode library 195

4.5 DISCUSSION OF CHAPTER 4 RESULTS 197

4.5.1 Barcode libraries of low complexity can be analysed using next generation sequencing but moderately complex barcode libraries cannot 197

4.5.2 Stochastic and systematic sequencing errors give rise to high frequency false barcodes and limit the specificity of analysis of complex barcode libraries 198

4.5.3 Limited sensitivity of next generation sequencing for detecting certain barcode sequences 200

4.5.4 Limited potential for quantitative analysis 200

4.5.5 A possible alternative barcode design 201
4.5.6 Possible implications for next generation sequencing analyses of TCR repertoire 201

4.5.7 Insights into sequencing error 202

4.6 CONCLUSIONS FOR CHAPTER 4 203

CHAPTER FIVE: GENE THERAPY RESEARCHERS’ ASSESSMENTS OF RISKS AND PERCEPTIONS OF RISK ACCEPTABILITY IN CLINICAL TRIALS 205

5.1 INTRODUCTION 205

5.2 CHAPTER 5 HYPOTHESIS AND AIMS 208

5.3 MATERIALS AND METHODS 209

5.3.1 Study design 209

5.3.2 Construction and validation of the survey instrument 213

5.3.3 Data analysis and interpretation 214

5.4 RESULTS 215

5.4.1 Demographics of respondents 215

5.4.2 Influence of perceptions of clinical context and attitudes towards preclinical evidence on assessments of risks in a hypothetical clinical trial 216

5.4.3 Differential risk assessment by demographic groups when making decisions about initiating a hypothetical clinical trial 221

5.4.4 Attitudes towards research involving human subjects 228

5.5 DISCUSSION OF CHAPTER 5 RESULTS 231

5.5.1 Study limitations 231

5.5.2 Influence of the clinical context for clinical trial decision-making 233

5.5.3 Influence of the strength and utility of preclinical evidence on clinical trial decision-making 234

5.5.4 Experiential influences on clinical trial decision-making 235

5.5.5 Impact of serious adverse events on clinical trial decision-making 236

5.6 CONCLUSIONS FOR CHAPTER 5 237
CHAPTER SIX: GENERAL DISCUSSION 239

6.1 INTRODUCTION 239

6.2 SUMMARY OF OUTCOMES OF THE BIOLOGICAL ANALYSES 240

6.2.1 Effect of promoter-enhancer transcriptional activity on clonal complexity and proliferative stress in reconstituted lymphocytes 240

6.2.2 Feasibility of vector barcoding and analysis of complex barcode libraries using next generation sequencing 242

6.3 LIMITATIONS OF TECHNOLOGIES, METHODOLOGIES AND MODEL SYSTEMS USED FOR PERFORMING BIOLOGICAL MEASUREMENTS AND PRECLINICAL ASSESSMENTS 243

6.3.1 Limitations of next generation sequencing technology 243

6.3.2 Limitations of integration site analysis methodologies 247

6.3.3 Limitations of small animal models 248

6.4 POSSIBLE TRANSLATIONAL UNCERTAINTIES ABOUT SAFETY AND EFFICACY THAT MAY REMAIN AFTER PRECLINICAL ANALYSES 250

6.5 THE LEVEL OF CERTAINTY REQUIRED FOR DECISION-MAKING 252

6.6 SUMMARY OF OUTCOMES OF THE ETHICAL ANALYSES 253

6.7 INFLUENCE OF CLINICAL CONTEXT ON BIOLOGICAL ANALYSES 253

6.8 INTERDEPENDENCE BETWEEN BIOLOGICAL AND ETHICAL ANALYSES IN THE INTERPRETATION OF RESULTS DESCRIBED IN THIS THESIS 256

6.9 CONCLUSIONS AND SIGNIFICANCE OF THIS WORK 258

References 260

Appendix 1: Customised Perl scripts used to analyse next generation sequencing data in Chapter 4 286

Appendix 2: Survey instrument used in Chapter 5 302

Appendix 3: Publications arising from this work 313
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full</th>
<th>Page first used</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAV</td>
<td>Adeno-associated virus</td>
<td>31</td>
</tr>
<tr>
<td>ADA</td>
<td>Adenosine deaminase</td>
<td>4</td>
</tr>
<tr>
<td>ADA-SCID</td>
<td>Severe combined immunodeficiency caused by adenosine deaminase</td>
<td>14</td>
</tr>
<tr>
<td>AGTS</td>
<td>Australasian Gene Therapy Society</td>
<td>212</td>
</tr>
<tr>
<td>ALD</td>
<td>Adrenoleukodystrophy</td>
<td>68</td>
</tr>
<tr>
<td>ALL</td>
<td>Acute lymphoblastic leukaemia</td>
<td>17</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance of groups</td>
<td>114</td>
</tr>
<tr>
<td>APC</td>
<td>Allophycocyanin</td>
<td>102</td>
</tr>
<tr>
<td>ARC</td>
<td>Animal Resources Centre</td>
<td>99</td>
</tr>
<tr>
<td>ASGCT</td>
<td>American Society of Gene and Cell Therapy</td>
<td>26</td>
</tr>
<tr>
<td>bp</td>
<td>Base pairs</td>
<td>52</td>
</tr>
<tr>
<td>BM</td>
<td>Bone marrow</td>
<td>91</td>
</tr>
<tr>
<td>BMT</td>
<td>Bone marrow transplantation</td>
<td>7</td>
</tr>
<tr>
<td>BSGCT</td>
<td>British Society for Gene and Cell Therapy</td>
<td>212</td>
</tr>
<tr>
<td>CAG</td>
<td>Cytomegalovirus enhancer/ chicken β-actin</td>
<td>58</td>
</tr>
<tr>
<td>CGD</td>
<td>Chronic granulomatous disease</td>
<td>24</td>
</tr>
<tr>
<td>cHS4</td>
<td>Chicken β-globin locus hypersensitive site 4 insulator</td>
<td>56</td>
</tr>
<tr>
<td>CMV</td>
<td>Cytomegalovirus</td>
<td>55</td>
</tr>
<tr>
<td>cPPT</td>
<td>Central polypurine tract</td>
<td>55</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl sulphoxide</td>
<td>76</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
<td>75</td>
</tr>
<tr>
<td>EF1α</td>
<td>Elongation factor-1-α</td>
<td>31</td>
</tr>
<tr>
<td>EFS</td>
<td>Intron-less (short) form of elongation factor-1-α</td>
<td>51</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>ESGCT</td>
<td>European Society for Gene and Cell Therapy</td>
<td>212</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
<td>102</td>
</tr>
<tr>
<td>FITC</td>
<td>Fluorescein isothiocyanate</td>
<td>102</td>
</tr>
<tr>
<td>FV</td>
<td>Foamyvirus</td>
<td>30</td>
</tr>
<tr>
<td>GALV</td>
<td>Gibbon ape leukaemia virus</td>
<td>13</td>
</tr>
<tr>
<td>GEEs</td>
<td>Generalised estimating equations</td>
<td>214</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
<td>23</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>Granulocyte-macrophage colony stimulating factor</td>
<td>26</td>
</tr>
<tr>
<td>GTRU</td>
<td>Gene Therapy Research Unit</td>
<td>89</td>
</tr>
<tr>
<td>GvHD</td>
<td>Graft-versus-host disease</td>
<td>7</td>
</tr>
<tr>
<td>H &amp; E</td>
<td>Haematoxylin and eosin</td>
<td>104</td>
</tr>
<tr>
<td>HBS</td>
<td>Hepes-buffered saline</td>
<td>96</td>
</tr>
<tr>
<td>HIV-1</td>
<td>Human immunodeficiency virus type-1</td>
<td>30</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
<td>7</td>
</tr>
<tr>
<td>HPC</td>
<td>Haematopoietic progenitor cell</td>
<td>7</td>
</tr>
<tr>
<td>IEC</td>
<td>Institutional ethics committee</td>
<td>32</td>
</tr>
<tr>
<td>Ig</td>
<td>Immunoglobulin</td>
<td>4</td>
</tr>
<tr>
<td>IL</td>
<td>Interleukin</td>
<td>4</td>
</tr>
<tr>
<td>IVIG</td>
<td>Intravenous immunoglobulin</td>
<td>17</td>
</tr>
<tr>
<td>JAK</td>
<td>Janus kinase</td>
<td>4</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobases</td>
<td>41</td>
</tr>
<tr>
<td>LAM-PCR</td>
<td>Linear amplification-mediated PCR</td>
<td>45</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani broth</td>
<td>76</td>
</tr>
<tr>
<td>LCA</td>
<td>Leber's congenital amaurosis</td>
<td>68</td>
</tr>
<tr>
<td>LMO2</td>
<td>LIM domain only 2</td>
<td>22</td>
</tr>
<tr>
<td>LM-PCR</td>
<td>Ligation-mediated PCR</td>
<td>45</td>
</tr>
<tr>
<td>LM-PCRI</td>
<td>Initial PCR amplification during ligation-mediated PCR</td>
<td>111</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Term</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>LTR</td>
<td>Long terminal repeat</td>
<td>12</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
<td>126</td>
</tr>
<tr>
<td>MFE</td>
<td>Minimum Gibbs Free Energy</td>
<td>192</td>
</tr>
<tr>
<td>MLV</td>
<td>Murine leukaemia virus</td>
<td>30</td>
</tr>
<tr>
<td>MoMLV</td>
<td>Moloney murine leukaemia virus</td>
<td>12</td>
</tr>
<tr>
<td>MSCV</td>
<td>Murine stem cell virus</td>
<td>31</td>
</tr>
<tr>
<td>MUD</td>
<td>Matched unrelated donor</td>
<td>7</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
<td>3</td>
</tr>
<tr>
<td>NK</td>
<td>Natural killer</td>
<td>4</td>
</tr>
<tr>
<td>nPCR</td>
<td>Nested PCR</td>
<td>111</td>
</tr>
<tr>
<td>nrLAM-PCR</td>
<td>Non-restrictive linear amplification-mediated PCR</td>
<td>47</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
<td>112</td>
</tr>
<tr>
<td>OTC</td>
<td>Ornithine transcarbamylase</td>
<td>68</td>
</tr>
<tr>
<td>PBS-/-</td>
<td>Phosphate buffered saline, without calcium and magnesium</td>
<td>75</td>
</tr>
<tr>
<td>PE</td>
<td>Phycoerythrin</td>
<td>102</td>
</tr>
<tr>
<td>PGK</td>
<td>Phosphoglycerate kinase</td>
<td>31</td>
</tr>
<tr>
<td>PNP</td>
<td>Purine nucleoside phosphorylase</td>
<td>4</td>
</tr>
<tr>
<td>Q30</td>
<td>Quality filtering based on a phred score of 30</td>
<td>166</td>
</tr>
<tr>
<td>QPL</td>
<td>Question prompt list</td>
<td>70</td>
</tr>
<tr>
<td>RAG1</td>
<td>Recombinase activating gene-1</td>
<td>4</td>
</tr>
<tr>
<td>RAG2</td>
<td>Recombinase activating gene-2</td>
<td>4</td>
</tr>
<tr>
<td>RCL</td>
<td>Replication-competent lentivirus</td>
<td>55</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust multiarray averaging</td>
<td>107</td>
</tr>
<tr>
<td>RRE</td>
<td>Rev response element</td>
<td>52</td>
</tr>
<tr>
<td>RSV</td>
<td>Rous sarcoma virus</td>
<td>55</td>
</tr>
<tr>
<td>SA</td>
<td>Splice acceptor</td>
<td>21</td>
</tr>
<tr>
<td>SAE</td>
<td>Serious adverse event</td>
<td>1</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>SCID-X1</td>
<td>X-linked severe combined immunodeficiency</td>
<td>1</td>
</tr>
<tr>
<td>SCID</td>
<td>Severe combined immunodeficiency</td>
<td>3</td>
</tr>
<tr>
<td>SD</td>
<td>Splice donor</td>
<td>21</td>
</tr>
<tr>
<td>SFFV</td>
<td>Spleen focus-forming virus</td>
<td>50</td>
</tr>
<tr>
<td>SFTCG</td>
<td>French Society of Cell and Gene Therapy</td>
<td>212</td>
</tr>
<tr>
<td>SIN</td>
<td>Self-inactivating</td>
<td>49</td>
</tr>
<tr>
<td>SIV</td>
<td>Simian immunodeficiency</td>
<td>44</td>
</tr>
<tr>
<td>SOC</td>
<td>Super optimal broth with carbolite repression</td>
<td>76</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Sequencing by oligonucleotide ligation and detection</td>
<td>146</td>
</tr>
<tr>
<td>SPF</td>
<td>Specific pathogen free</td>
<td>99</td>
</tr>
<tr>
<td>SURE</td>
<td>Stop unwanted recombination events</td>
<td>76</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate ethylenediaminetetraacetic acid</td>
<td>82</td>
</tr>
<tr>
<td>TALEN</td>
<td>Transcription activator-like effector nuclease</td>
<td>62</td>
</tr>
<tr>
<td>T-ALL</td>
<td>T cell acute lymphoblastic leukaemia</td>
<td>20</td>
</tr>
<tr>
<td>TBE</td>
<td>Tris-borate ethylenediaminetetraacetic acid</td>
<td>82</td>
</tr>
<tr>
<td>TCR</td>
<td>T cell receptor</td>
<td>4</td>
</tr>
<tr>
<td>TE</td>
<td>Tris-ethylenediaminetetraacetic acid</td>
<td>79</td>
</tr>
<tr>
<td>TEMED</td>
<td>N,N,N',N'-tetramethylethlenediamine</td>
<td>83</td>
</tr>
<tr>
<td>TK</td>
<td>Thymidine kinase</td>
<td>57</td>
</tr>
<tr>
<td>TU</td>
<td>Transducing units</td>
<td>97</td>
</tr>
<tr>
<td>U3</td>
<td>Unique 3</td>
<td>20</td>
</tr>
<tr>
<td>UCOE</td>
<td>Ubiquitously acting chromatin opening element</td>
<td>58</td>
</tr>
<tr>
<td>VCN</td>
<td>Vector copy number</td>
<td>104</td>
</tr>
<tr>
<td>VSVG</td>
<td>Vesicular stomatitis virus G glycoprotein</td>
<td>55</td>
</tr>
<tr>
<td>WAS</td>
<td>Wiskott-Aldrich syndrome</td>
<td>24</td>
</tr>
<tr>
<td>WASP</td>
<td>Wiskott-Aldrich syndrome protein</td>
<td>89</td>
</tr>
</tbody>
</table>
# TABLE OF FIGURES

| Figure 1.1 | The common gamma chain of the interleukin receptors 2, 4, 7, 9, 15 and 21, and the downstream signal transduction pathway | 5 |
| Figure 1.2 | Requirement for γc-dependent interleukin signalling to enable lymphocyte ontogeny during haematopoiesis | 6 |
| Figure 1.3 | The interdependence between biological and ethical analyses of safety and efficacy in translational gene therapy | 63 |
| Figure 3.1 | Schematic diagram of mouse model for SCID-X1 gene therapy | 101 |
| Figure 3.2 | Methodology for preparing libraries of LTR-genomic DNA fragments for integration site analysis | 108 |
| Figure 3.3 | Structure of integration site libraries and bioinformatic workflow for analysis of sequence data | 113 |
| Figure 3.4 | Schematic diagram of self-inactivating lentiviral vectors EF1α.γc and PGK.γc, and relative strengths of the EF1α and PGK promoter-enhancer elements used | 116 |
| Figure 3.5 | Levels of reconstituted lymphocyte subsets at 6 weeks post-transplantation in γc-deficient mice treated with EF1α.γc (n = 27) and PGK.γc (n = 26) lentiviral vectors, or a wild-type transplant (n = 27) | 117 |
| Figure 3.6 | Levels of reconstituted lymphocyte subsets at 16 weeks post-transplantation in γc-deficient mice treated with EF1α.γc (n = 28) and PGK.γc (n = 28) lentiviral vectors, or a wild-type transplant (n = 27) | 119 |
| Figure 3.7 | Ratios of CD3⁺ T cells to B220⁺ B cells, and of CD4⁺ T cells to CD8⁺ T cells in mice treated with EF1α.γc, PGK.γc and wild-type Sca1⁺ bone marrow cells | 121 |
| Figure 3.8 | Proportion of lymphoma-free survival in mice treated with EF1α.γc or PGK.γc lentiviral vectors, or wild-type Sca1⁺ bone marrow cells | 122 |
| Figure 3.9 | Lymphoma infiltration of tissue | 125 |
Figure 3.10  Microarray analysis of relative expression of genes within 1 Mb lentiviral vector integration in lymphoma samples from mice treated with EF1α.γc or PGK.γc lentiviral vectors

Figure 4.1  Schematic diagram of primer extension method used to construct a barcode insert

Figure 4.2  Validation of barcode insert construction and preliminary validation of library complexity

Figure 4.3  Analysis of a sample containing 2 defined barcode sequences

Figure 4.4  Abundance of unique barcodes detected by Illumina next generation sequencing in a sample containing only 2 known barcodes, after Q30 quality filtering and clustering with 1 mismatch

Figure 4.5  Abundance of unique barcodes detected by Illumina next generation sequencing analysis of highly complex samples

Figure 4.6  Analyses of the distribution of error frequencies and read qualities for each position in sequence reads

Figure 4.7  Schematic diagram of primer extension method used to construct a barcode insert containing adaptor sequences specific for sequencing on the Illumina platform

Figure 4.8  Schematic diagram of primer extension method used to construct a barcode insert containing adaptor sequences specific for sequencing on the SOLiD platform

Figure 4.9  PCR amplification to introduce adaptor sequences and sample indices to Illumina-specific and SOLiD-specific barcodes

Figure 4.10  Distribution of the relative abundance of the 500 most abundant barcode sequences detected following analysis of the defined barcode libraries using different sequencing platforms

Figure 4.11  qPCR confirmation of the absence of false barcodes produced by sequencing error in the Illumina-specific 100-barcode library

Figure 4.12  Analysis of the relative abundance, GC content and likelihood of secondary structure formation for each of the 100 expected Illumina-compatible barcode sequences
Figure 4.13  qPCR confirmation of the presence of selected expected barcodes in the Illumina-specific 100-barcode library

Figure 4.14  Analysis of the position and substitution-like type of all 1-mismatch sequence errors for both Illumina HiSeq2000 sequencing runs

Figure 5.1  Percentage of respondents who agreed it would be appropriate to recruit subjects to a phase I clinical trial for each of the hypothetical clinical scenarios based on data generated in each of the preclinical models

Figure 5.2  Consistent patterns of odds ratios when separate GEE models were fitted for each of the preclinical models

Figure 5.3  Consistent patterns of odds ratios when separate GEE models were fitted for each of the clinical scenarios

Figure 5.4  Heat-map representing the proportions of respondents in each demographic group who supported initiation of a clinical trial, for each of the clinical scenarios and types of preclinical models
### TABLE OF TABLES

| Table 1.1 | Long-term outcomes of bone marrow transplantation in patients with severe combined immunodeficiency | 9 |
| Table 1.2 | Atypical SCID-X1 phenotypes and implications for SCID-X1 gene therapy | 11 |
| Table 1.3 | Trials of gene therapy targeting other diseases affecting the haematopoietic system | 15 |
| Table 1.4 | Long-term outcomes of clinical trials of gene therapy for SCID-X1 | 16 |
| Table 1.5 | Analysis of leukaemias in gene therapy trials | 18-19 |
| Table 1.6 | Summary of key findings of studies investigating retroviral integration behaviour and retroviral-induced clonal dominance | 35-40 |
| Table 1.7 | Summary of functional characteristics of HIV-1 wild type sequences and use in third generation lentiviral vectors | 53-54 |
| Table 2.1 | Buffers and solutions used for tissue culture | 73-75 |
| Table 2.2 | Genotypes of bacterial strains | 77 |
| Table 2.3 | Liquid and solid media used for bacterial work | 77 |
| Table 2.4 | Buffers and solutions used for molecular biology | 80 |
| Table 3.1 | Plasmids used for lentiviral vector production | 95 |
| Table 3.2 | Oligonucleotide sequences used for assigning titre to lentiviral vector stocks | 98 |
| Table 3.3 | Antibodies used in flow cytometry analysis of immune reconstitution | 103 |
| Table 3.4 | Oligonucleotide sequences used for conventional LM-PCR | 105 |
| Table 3.5 | Oligonucleotide sequences used for preparation of vector-genomic junction fragment libraries | 110 |
Table 3.6 Macroscopic and cellular characteristics of lymphomas arising in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

Table 3.7 Characteristics of vector integrations in lymphomas arising in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

Table 3.8 Numbers and proportions of sequence reads for each sample that were retained after each stage of data processing

Table 3.9 Total mapped sequence reads and numbers of unique integration sites

Table 4.1 Oligonucleotide sequences used for construction of barcode inserts

Table 4.2 Oligonucleotide sequences used for amplification of barcodes as preparation for next generation sequencing

Table 4.3 Oligonucleotide sequences used for validation of the Illumina-specific 100-barcode library

Table 4.4 Effect of bioinformatic strategies on the reduction of background detected in a sample containing 2 known barcodes

Table 4.5 Analysis of unique barcodes detected in samples representing the complex plasmid library and vector stock

Table 4.6 Effect of analytical strategies on the reduction of background caused by sequencing error for the Illumina-specific barcode libraries of defined complexity

Table 4.7 Relative abundance of specific barcode sequences within the 100-barcode library as measured by the Illumina HiSeq2000 and quantitative PCR

Table 4.8 Effect of analytical strategies on the reduction of background caused by sequencing error for the SOLiD-specific barcode libraries of defined complexity

Table 4.9 Number of unique barcodes detected in the highly complex Illumina-specific barcode library before and after quality filtering
| Table 4.10 | Number of unique barcodes detected in serial dilutions of the highly complex Illumina-specific barcode library | 196 |
| Table 5.1 | Descriptions of preclinical evidence provided to survey participants for decision-making about the initiation of a clinical trial | 210 |
| Table 5.2 | Clinical scenarios provided to survey participants for decision-making about the initiation of a clinical trial | 211 |
| Table 5.3 | Demographic details of survey respondents | 216 |
| Table 5.4 | Analysis of the effects of varying clinical scenarios for GEE models fitted individually for each of the types of preclinical evidence | 220 |
| Table 5.5 | Analysis of the effects of varying preclinical evidence for GEE models fitted individually for each clinical scenario | 223 |
| Table 5.6 | Effects of professional experience on decision-making in the least severe clinical scenarios when preclinical evidence involved cell culture-based models or there was a related phase I study with low frequency adverse events | 226-227 |
| Table 5.7 | Percentage of respondents who agreed with statements about attitudes towards research involving human subjects | 230 |
Gene therapy, the introduction of nucleic acids (RNA or DNA) or synthetic analogues into somatic cells to treat human disease, is a promising and relatively new medical field. A major milestone in the history of gene therapy was achieved in 2000 with the report that gene therapy had reconstituted immune function in 2 infants with X-linked severe combined immunodeficiency (SCID-X1; Cavazzana-Calvo et al. 2000). However the development of leukaemia in 5 of 20 infants treated worldwide overshadowed this initial success, because leukaemogenesis was directly linked to the integrating viral vector used for gene delivery. These serious adverse events (SAEs) have prompted much research into designing safer vectors while retaining efficacy, but also highlight the difficulties associated with the uncertain nature of risk in translational research.
In this research project, a novel, interdisciplinary approach is taken to address the challenge of analysing safety and efficacy in translational gene therapy. This introduction articulates how the different disciplines fit together in this project, by identifying safety and efficacy as common threads running through this thesis. The development of gene therapy for SCID-X1 is outlined, including the genotoxic risks underlying the leukaemias in the SCID-X1 trials. Strategies for improving the safety of SCID-X1 gene therapy are outlined. Integration site methodologies for quantifying the number of haematopoietic progenitor cells giving rise to mature lymphocytes following gene therapy, used for monitoring safety and efficacy, are reviewed. Finally, a philosophical consideration of the assessment of uncertain risks in translational gene therapy is presented.

### 1.1 SAFETY AND EFFICACY IN TRANSLATIONAL GENE THERAPY

A cornerstone of translational research is that any novel therapeutic product, such as a gene transfer vector, must be shown to be safe and efficacious using the preclinical model systems that best recapitulate human disease before a phase I human trial can be contemplated. Measurements of safety and efficacy in preclinical models, however, have limited potential for predicting outcomes in human trials. For example, preclinical testing of the gene transfer vectors using cell culture and mouse models did not predict the frequency at which leukaemias developed in the SCID-X1 gene therapy trials (Section 1.4.1).

The manner in which safety and efficacy are analysed at the preclinical stage of translational research is the unifying theme of this research project. The biological component of this project uses a mouse model of SCID-X1 gene therapy to characterise
Chapter 1: Introduction

The safety and efficacy profile of novel gene transfer vectors, designed in response to the SAEs in the SCID-X1 trials. In addition, this project has developed barcoded vectors as tools for analysing the safety and efficacy of SCID-X1 gene therapy. The feasibility of analysing complex barcoded plasmid libraries using next generation sequencing (NGS) is evaluated. Using gene therapy for SCID-X1 as an exemplar for the uncertainty of risk in clinical research, the ethical component of this project stems from the notion that even when the best available model systems and technologies are used to analyse the safety and efficacy of a novel therapy at the preclinical stage, a degree of uncertainty is inescapable when human clinical trials are initiated. This project has investigated how gene therapy researchers perceive and assess risks when designing and conducting clinical trials.

The interdisciplinary nature of this research project reflects the interdisciplinary nature of the problems requiring consideration in translational medicine. By evaluating issues of safety and efficacy associated translational gene therapy at the stages of generating preclinical evidence and planning clinical trials, and highlighting the interaction between biological and ethical analyses of safety and efficacy, this work represents a novel contribution to the development of translational SCID-X1 gene therapy.

1.2 DEVELOPMENT OF GENE THERAPY FOR X-LINKED SEVERE COMBINED IMMUNODEFICIENCY (SCID-X1)

1.2.1 Background on SCID-X1

Severe combined immunodeficiencies (SCIDs) are inherited disorders in which a genetic defect causes failure of the development, maturation, survival or function of
immune lymphocytes, including T cells, B cells and natural killer (NK) cells. Severely affected infants present clinically with persistent infections, failure to thrive and, in the absence of treatment, die within the first year of life (Buckley 2004). SCIDs can be caused by mutations affecting cytokine receptors such as interleukin (IL) receptors (Puck et al. 1993; Noguchi et al. 1993); the tyrosine phosphatase CD45 (Kung et al. 2000); the γ, δ or ε invariable chains of CD3 associated with the T cell receptor (TCR; Arnaiz-Villena et al. 1993; Dadi et al. 2003; de Saint Basile et al. 2004); cytokine signalling molecules such as Janus kinase 3 (JAK3; Russell et al. 1995); enzymes such as adenosine deaminase (ADA) and purine nucleoside phosphorylase (PNP; Giblett et al. 1972; Giblett et al. 1975); and recombinase activating gene-1 (RAG1), recombinase activating gene-2 (RAG2) and Artemis, which are involved in V(D)J recombination and non-homologous end-joining during DNA repair (Schwarz et al. 1996; Moshous et al. 2001).

The X-linked form of SCID accounts for about 70% of inherited SCIDs and affects approximately 1 in 50,000 to 1 in 100,000 live births (Puck et al. 1997; Buckley 2002; Fischer 2004). SCID-X1 is caused by mutations in \( IL2R\gamma \), originally cloned as a component of the IL-2 receptor (Noguchi et al. 1993; Puck et al. 1993). \( IL2R\gamma \) encodes the \( \gamma c \) protein, the common gamma chain of the receptors for IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21 (Figure 1.1). Mutations in \( IL2R\gamma \) block growth signalling via these IL receptors, such that failed IL-7 signalling blocks T cell ontogeny and failed IL-15 signalling blocks NK cell ontogeny (Figure 1.2) (Cavazzana-Calvo et al. 1996; Puel et al. 1998; Kennedy et al. 2000). While B cell numbers remain relatively normal, B cell function is defective because T cell help is absent, resulting in decreased serum immunoglobulin (Ig) levels and subsequently, an increased risk of infection.
Figure 1.1: The common gamma chain of the interleukin receptors 2, 4, 7, 9, 15 and 21, and the downstream signal transduction pathway. The common gamma chain (γc) subunit of the interleukin (IL) receptors 2, 4, 7, 9, 15 and 21 is shown in red. The extracellular domain of each receptor binds the appropriate IL ligand, permitting the intracellular domain of γc to bind Janus kinase 3 (JAK3), which is activated by autophosphorylation. Activated JAK3 phosphorylates signal transducer and activator of transcription-5 (STAT5), which dimerises and enters the nucleus. Phosphorylated STAT5 activates genes required for cell growth. Mutations in γc block or reduce growth signalling through these IL receptors.
Figure 1.2: Requirement for γc-dependent interleukin signalling to enable lymphocyte ontogeny during haematopoiesis. Haematopoietic stem cells (HSCs) give rise to common lymphoid progenitor cells (LPCs), and ultimately B cells, CD4+ T cells, CD8+ T cells and NK cells after many rounds of cell division and differentiation. Mutations in γc block IL-7 and IL-15 signalling, essential for T and NK cell ontogeny, respectively. Correction of a γc mutation in low numbers of HSC or LPC, or transplantation of wild type HSC and LPC, can restore T and NK cell ontogeny, because these haematopoietic progenitors undergo massive replicative expansion. Hence HSCs and LPCs transduced with γc have a selective advantage in gene therapy.
1.2.2 Limitations of existing treatment for SCID-X1 and rationale for development of gene therapy

In 1968 the first successful bone marrow transplantation (BMT) treated 2 patients, including 1 SCID patient, using bone marrow from a related, human leukocyte antigen (HLA)-identical donor (Gatti et al. 1968; Bach et al. 1968). Conventional treatment for SCID patients since has comprised an allogeneic BMT, ideally from an HLA-identical sibling donor. Following engraftment, haematopoietic progenitor cells (HPCs) expand and differentiate into T cells and NK cells (Figure 1.2). This treatment has a centre-dependent survival rate of over 90% and remains the best treatment for SCID-X1 when available. Unfortunately, less than a third of SCID-X1 patients have access to an HLA-identical donor (Cavazzana-Calvo et al. 2004). In Australia, only 28 of 135 patients (20.7%) with primary immunodeficiencies transplanted between 1992-2008 had access to an HLA-identical sibling donor (Mitchell et al. 2013).

For patients lacking an HLA-identical sibling donor, BMT from a haplo-identical parent or matched unrelated donor (MUD) represents the only available curative treatment for SCID-X1. In each case this is associated with a high degree of risk of death due to graft-versus-host disease (GvHD). Donor availability is also a limitation for unrelated BMT, as even with approximately 21 million donors registered worldwide there remains an under-representation of non-Caucasian ethnic groups, meaning that the possibility of identifying a suitably-matched donor is much lower for patients whose ethnic ancestry is not north Caucasian. Two groups of BMT centres and 1 individual BMT centre have published long-term survival outcomes for over 100 SCID patients or patients with primary immunodeficiencies (Buckley, 2011; Gennery et al. 2010; Haddad et al. 1998; Mitchell et al. 2012). These results show that HLA-
mismatched BMT has a lower centre-dependent survival rate of 70%–80% at 5 years (Table 1.1). The major complications of BMT, many of which occur more commonly with unrelated donors, include infection, interstitial pneumonia syndromes, mucositis, sinusoidal obstructive syndrome, and particularly GvHD. GvHD occurs following a BMT, when contaminating T cells are present in transplanted bone marrow, and so bypass thymic elimination of auto-reactive T cells in the BMT recipient (Ferrara et al. 2009). These T cells identify the recipient’s tissue as foreign, resulting in an immune response directed against the skin, liver and gut in particular. Clinical outcomes following BMT from HLA-mismatched donors can also be limited by the older age of the recipient at transplantation, pre-existing infections, long-term decline in T cell function, slow or partial immune reconstitution, and GvHD (Table 1.1; Cavazzana-Calvo & Fischer 2007).

While HLA-identical BMT remains the treatment of choice for SCID-X1 patients, there is a clear need to develop alternative therapies, since the majority of SCID-X1 patients lack access to HLA-identical sibling donors and mismatched BMT is associated with a high risk of mortality and morbidity. Gene therapy is a potentially attractive alternative treatment for SCID-X1. Delivery of IL2Rγ may restore γc expression and γc-dependent signalling in the patient’s own HPCs. This may enable those genetically modified HPCs to respond to γc-dependent signals, and expand and differentiate into mature T and NK cells and correct B cell function, thus restoring cellular and humoral immunity without the risk of GvHD.
Table 1.1: Long-term outcomes of bone marrow transplantation in patients with severe combined immunodeficiency

<table>
<thead>
<tr>
<th>Location of centre(s)</th>
<th>Disease treated</th>
<th>Dates of BMT</th>
<th>Number of patients treated</th>
<th>Number of patients with T cell engraftment (percentage)</th>
<th>Short-term BMT-related morbidity rate</th>
<th>Long-term survival rate</th>
<th>Survival of recipients of related HLA-identical bone marrow</th>
<th>Factors associated with survival</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian and New Zealand Children’s Haematology Oncology Group transplantation centres</td>
<td>All primary IDs</td>
<td>1992–2008</td>
<td>135</td>
<td>NR</td>
<td>10% at day 100, 22% at 1 year</td>
<td>72% at 5 years post-BMT (70% for SCID patients)</td>
<td>79% survival at 5 years post-BMT (n = 38)</td>
<td>Absence of pre-existing infections</td>
<td>Mitchell et al. 2013</td>
</tr>
<tr>
<td>37 European BMT centres</td>
<td>All primary IDs</td>
<td>1968–2005</td>
<td>699</td>
<td>NR</td>
<td>NR</td>
<td>71% survival at 10 years post-BMT (n = 135)</td>
<td>84% survival at 10 years post-BMT</td>
<td>B- phenotype, absence of pre-existing infections, liver impairment and malnutrition</td>
<td>Gennery et al. 2010</td>
</tr>
<tr>
<td>18 European BMT centres</td>
<td>SCID</td>
<td>1983–1993</td>
<td>193</td>
<td>116 (at 6 months post-BMT)</td>
<td>39.9% at 6 months</td>
<td>NR</td>
<td>Not applicable: patients received HLA-non-identical bone marrow</td>
<td>Absence of T cell reconstitution at 6 months post-BMT</td>
<td>Haddad et al. 1998</td>
</tr>
<tr>
<td>Duke University Medical Center</td>
<td>SCID (45% had SCID-X1)</td>
<td>1982–2010</td>
<td>166</td>
<td>121</td>
<td>24%</td>
<td>75% at 1 year post-BMT, 66% at 5 years post-BMT</td>
<td>100% survival of recipients of HLA-identical bone marrow (n = 17)</td>
<td>BMT performed within first 3.5 months of life</td>
<td>Buckley 2011</td>
</tr>
</tbody>
</table>

BMT, bone marrow transplant; ID, immunodeficiency; NR, not reported; SCID, severe combined immunodeficiency
1.2.3 Lessons for SCID-X1 gene therapy from spontaneous reversion events and partial SCID-X1 phenotypes

Spontaneous mutations have occurred in several patients with immunodeficiencies affecting T cell development, permitting a single HPC to revert to either a wild-type phenotype or a phenotype permitting differentiation into functional T cells (Stephan et al. 1996; Hirschhorn et al. 1996; Speckmann et al. 2008). Analysis of a patient with a partial SCID-X1 phenotype revealed a single HPC had reverted to wild type, divided 10–11 times, with the resulting cell population comprising mature T cells with unique TCR Vβ sequences (Bousso et al. 2000). These clones expanded to approximately 50% of normal T cell numbers, with an in vivo ability to defend against infections. Not only do these reversion events demonstrate the proliferative capacity of HPCs and the long-term persistence of mature T cells, they illustrate the therapeutic potential that may be achieved by genetically repairing low numbers of HPCs to express γc.

While the classical SCID-X1 phenotype is T⁻NK⁻, atypical T⁻NK⁺ and T⁺NK⁺ phenotypes have also been reported, with approximately 10% of SCID-X1 patients having a T⁻NK⁺ phenotype. These atypical phenotypes result from mutations affecting γc binding affinities and IL2Rγ splice sites, effectively reducing functional levels of γc (Table 1.2). Useful lessons for gene therapy can be derived from these atypical phenotypes. Firstly, there is a threshold of γc expression required for signal transduction via both the IL-7R and IL-15R, and hence for both T cell and NK cell ontogeny (Smyth et al. 2007). Secondly, limiting γc expression is more likely to impair T cell reconstitution, with NK cell ontogeny preferentially retained as γc becomes limiting. Thirdly, even if γc expression is adequate to enable both T cell and NK cell ontogeny, relatively low γc expression levels may impair the diversity of the TCR
Table 1.2: Atypical SCID-X1 phenotypes and implications for SCID-X1 gene therapy

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Location of mutation</th>
<th>Effect on γc expression levels</th>
<th>Effect on γc-dependent signalling</th>
<th>Lesson for SCID-X1 gene therapy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+NK+; T cells have poor function and oligoclonal T cell receptor diversity</td>
<td>G to A substitution at bp 115 at the first exon-intron boundary, within splice site</td>
<td>A more abundant aberrantly spliced non-functional γc mRNA variant and a less abundant correctly-sized γc mRNA</td>
<td>Level of high-affinity IL-2 receptors was reduced by 4-5-fold compared to healthy controls</td>
<td>If γc expression is lower than physiological levels, T cells and NK cells may develop but function and diversity of T cells may be reduced</td>
<td>DiSanto et al. 1994</td>
</tr>
<tr>
<td>T+NK+; T cells have poor function but repertoire was not severely restricted</td>
<td>G to A substitution at the fifth base of intron 1, within splice site</td>
<td>Incorrectly spliced γc mRNA and low levels of correctly spliced γc mRNA</td>
<td>Not studied</td>
<td>If γc expression is lower than physiological levels, T cells and NK cells may develop but T cell function and T cell receptor repertoire cells may be reduced</td>
<td>Wada et al. 2008</td>
</tr>
<tr>
<td>T+NK+</td>
<td>A to C substitution at the third base of intron 3, within splice site</td>
<td>Two aberrantly spliced γc mRNA variants and trace levels of correctly spliced γc mRNA</td>
<td>Signal transduction was possible via the IL-15 receptor, required for NK cell ontogeny. Signal transduction was not possible via the IL-7 receptor, required for T cell ontogeny</td>
<td>If γc expression is further reduced, the degree of T cell reconstitution may be reduced relative to the degree of NK cell reconstitution</td>
<td>Ginn et al. 2004; Smyth et al. 2007</td>
</tr>
<tr>
<td>T+NK+; T cell counts within normal range, but diversity was oligoclonal. Reduced serum IgG and IgA</td>
<td>T485 to G substitution, resulting in a leucine 162 to arginine substitution</td>
<td>γc expression undetectable on cell surface</td>
<td>Phosphorylated STAT5 and JAK3 could not be detected</td>
<td></td>
<td>Mella et al. 2000</td>
</tr>
<tr>
<td>T+NK+; T cell counts within normal range</td>
<td>C678 to T substitution, resulting in an arginine 222 to cysteine substitution</td>
<td>γc expression undetectable on cell surface</td>
<td>Phosphorylated STAT5 and JAK3 could not be detected</td>
<td></td>
<td>Mella et al. 2000</td>
</tr>
<tr>
<td>T+NK+, normal T cell count</td>
<td>C664 to T substitution, resulting in an arginine to cysteine substitution, within the extracellular region of γc near the transmembrane domain</td>
<td>Normal levels of γc expression</td>
<td>Reduced ability to bind IL-2 and respond to antigen stimulation</td>
<td></td>
<td>Sharfe et al. 1997</td>
</tr>
</tbody>
</table>
repertoire of reconstituted T cells, and thus a patient’s ability to respond to a diversity of antigens.

**1.2.4 Strategy for SCID-X1 gene therapy**

The strategy for SCID-X1 gene therapy is to harvest autologous bone marrow from the patient, select for cells expressing CD34, a marker for HPCs, and then transduce these CD34⁺ cells *ex vivo* with an integrating vector encoding $IL2R\gamma$. Crucial to the success of this strategy is the selective advantage conferred on transduced cells receiving $\gamma c$ (Figure 1.2). Once re-infused into the patient, transduced cells can respond to $\gamma c$-dependent IL growth signals, expand and differentiate into mature T and NK cell populations. Moreover, $\gamma c$ expression is stable and is maintained when the transduced cells divide and differentiate, due to genomic integration of the vector. This strategy has been employed to treat in excess of 20 patients worldwide, and similar strategies have since been employed to treat other diseases affecting the haematopoietic compartment (Section 1.2.5).

Given that HPCs undergo multiple rounds of division during immune reconstitution, there is an absolute requirement for vectors capable of giving rise to transgene expression in progeny cells via, for example, vector integration into genomic DNA. Early gene delivery systems were thus based on gammaretroviruses such as the Moloney murine leukaemia virus (MoMLV), which became possible following the sequencing of the MoMLV genome and characterisation of its replication cycle (Cavazzana-Calvo & Fischer 2007). These vectors were rendered replication incompetent by replacing viral protein-encoding sequences with the therapeutic transgene. Only the long terminal repeats (LTRs), packaging signal and sequences
necessary for reverse transcription of the RNA genome and integration of the provirus were retained.

This strategy of gammaretroviral gene transfer was successfully employed by a French team led by Alain Fischer, who reported in 2000 successful use of the *IL2Rγ*-encoding MoMLV-based vector, MFG-B2γc (Hacein-Bey *et al.* 1996), to treat 2 SCID-X1 infants (Cavazzana-Calvo *et al.* 2000). The selective growth advantage conferred on γc⁺CD34⁺ cells overcame the inefficiency of gene transfer observed in earlier unsuccessful gene therapy clinical trials (Cavazzana-Calvo *et al.* 2000). A British team led by Adrian Thrasher also reported successful gene therapy for SCID-X1 in 4 infants using a similar vector and protocol, but with 3 differences (Gaspar *et al.* 2004). The British MoMLV-based MFG-γc vector was pseudotyped with a heterologous envelope protein from the gibbon ape leukaemia virus (GALV), and transduction conditions excluded foetal calf serum and used a lower concentration of IL-3 (20 μg/L). The outcomes are described in Section 1.3.

1.2.5 Trials of haematopoietic gene therapy targeting other diseases

The haematopoietic compartment is a particularly promising target for gene therapy, because HPCs are accessible, can be manipulated *ex vivo*, have capacity for self-renewal, and, together with progeny cells, can retain function over an individual’s lifetime. The possibility of isolating, manipulating and transducing CD34⁺ HPCs *ex vivo* overcomes the challenge of specific gene delivery to the correct target cell population, which remains a major hurdle for many other applications of gene therapy. Achieving physiological levels of therapeutic transgene expression in the correct tissue type is more challenging for *in vivo* gene delivery strategies. Such strategies must
ensure a systemically-administered vector can transduce sufficient cells of the correct type, and regulatory elements must be manipulated to restrict transgene expression to the correct tissue.

Following the success of the French and British SCID-X1 trials that used ex vivo strategies, similar ex vivo gene therapy strategies have been reported in multiple trials of gene therapy for other diseases affecting the haematopoietic compartment (Table 1.3). Most notably, therapeutic efficacy has been reported for the majority of patients with adenosine deaminase deficiency-SCID (ADA-SCID) treated in independent Italian and British trials, with no reports of SAEs to date (Aiuti et al. 2009; Gaspar, Cooray, Gilmour, Parsley, Zhang et al. 2011).

1.3 CLINICAL TRIALS OF GENE THERAPY FOR SCID-X1

1.3.1 Efficacy of SCID-X1 gene therapy in the French and British clinical trials

The most recent report of the French trial of SCID-X1 gene therapy describes long-term outcomes in 9 patients, who lacked an HLA-identical donor and were treated between 1999 and 2002 (Table 1.4; Hacein-Bey-Abina et al. 2010). Reconstituted T cells were detected in 8 patients, and sustained T cell reconstitution was reported for 7 patients. For those 7 patients, T cell counts were within the normal range for 6 patients and slightly below normal for 1 patient at 7.3–10.7 years after treatment. T cell reconstitution appeared to correlate with the dose of CD34+γc+ cells received by patients. CD8+ T cell counts were within the normal range for all patients, while CD4+ T cell counts were within the normal range for 4 patients and slightly below normal for 3 patients. A diverse range of TCR Vβ families were detected, indicating a diverse TCR
Table 1.3: Trials of gene therapy targeting other diseases affecting the haematopoietic system

<table>
<thead>
<tr>
<th>Disease</th>
<th>Type of vector</th>
<th>Number of patients treated</th>
<th>Measurements of therapeutic efficacy</th>
<th>Adverse events reported (years of follow-up)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA-SCID</td>
<td>Gammaretroviral</td>
<td>10</td>
<td>Lymphoid reconstitution, no need for ER (n = 8) or IVIG (n = 9) therapies</td>
<td>None (1.8–8.0)</td>
<td>Aiuti et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Gammaretroviral</td>
<td>6</td>
<td>Lymphoid reconstitution (n = 4), no need for ER (n = 3) or IVIG (n = 3) therapies</td>
<td>None (2.0–7.0)</td>
<td>Gaspar, Cooray, Gilmour, Parsley, Zhang et al. 2011</td>
</tr>
<tr>
<td>ALD</td>
<td>Lentiviral</td>
<td>2</td>
<td>Gene-marking of multiple lineages, progressive demyelination ceased</td>
<td>None (2.0–2.5)</td>
<td>Cartier et al. 2009</td>
</tr>
<tr>
<td>WAS</td>
<td>Gammaretroviral</td>
<td>2</td>
<td>Functional correction of lymphocytes and monocytes, clinical improvement</td>
<td>None at 2.5 months (2010 publication); leukaemia development in 1 out of 8 subsequent patients (1.3 years post-gene therapy)</td>
<td>Boztug et al. 2010; Braun et al. 2011</td>
</tr>
<tr>
<td>CGD</td>
<td>Gammaretroviral</td>
<td>2</td>
<td>Initial functional correction of phagocytes and clinical improvement, subsequent transgene silencing</td>
<td>Myelodysplasia in Patient 1 and multilineage dysplasia in Patient 2 (2.3 and 3.6 years post-gene therapy, respectively)</td>
<td>Ott et al. 2006; Stein et al. 2010</td>
</tr>
<tr>
<td>β-thalassemia</td>
<td>Lentiviral</td>
<td>1</td>
<td>Transfusion independence for 1.8 years</td>
<td>None, although dominance of 1 myeloid clone (2.8)</td>
<td>Cavazzana-Calvo et al. 2010</td>
</tr>
<tr>
<td>B-CLL</td>
<td>Lentiviral</td>
<td>3</td>
<td>CD19-specific immune response (n = 3), complete remission (n = 2)</td>
<td>None (0.5)</td>
<td>Kalos et al. 2011</td>
</tr>
</tbody>
</table>

ADA-SCID, adenosine deaminase-deficient severe combined immunodeficiency; ALD, adrenoleukodystrophy; B-ALL, B cell chronic lymphoblastic leukaemia; CGD, chronic granulomatous disease; ER, enzyme replacement; IVIG, intravenous immunoglobulin

*Data in this table is based on the most recent trial publication and may not necessarily be up to date, with the exception of the report of leukaemia development in 1 patient treated in the WAS trial.
### Table 1.4: Long-term outcomes of clinical trials of gene therapy for SCID-X1

<table>
<thead>
<tr>
<th>Trial</th>
<th>Patient</th>
<th>Age at treatment (months)</th>
<th>Follow-up (years)</th>
<th>CD34⁺γc⁺ cell dose (×10⁶ per kg body weight)</th>
<th>CD3⁺ T cell count (per μL)</th>
<th>Natural killer cells (per μL)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>French</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Patient 1</td>
<td>11</td>
<td>10.6</td>
<td>3</td>
<td>800</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Patient 2</td>
<td>8</td>
<td>10.7</td>
<td>5</td>
<td>1,494</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Patient 5</td>
<td>3</td>
<td>9.4</td>
<td>20</td>
<td>2,175</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Patient 6</td>
<td>6</td>
<td>8.5</td>
<td>1</td>
<td>1248</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Patient 7</td>
<td>11</td>
<td>8.8</td>
<td>4</td>
<td>1,368</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Patient 8</td>
<td>6</td>
<td>8.2</td>
<td>22</td>
<td>2,250</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Patient 10</td>
<td>9</td>
<td>7.3</td>
<td>11</td>
<td>2,573</td>
<td>31</td>
</tr>
<tr>
<td><strong>Reference range</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,140–2,812</td>
<td>57–814</td>
</tr>
</tbody>
</table>

| **British**<sup>b</sup> | Patient 1 | 10 | 8.9 | 6 | 1,950 | 50 |
|                       | Patient 2 | 10 | 8.4 | 5 | 1,110 | 450 |
|                       | Patient 3 | 4 | 8.3 | 8 | 1,754 | 93 |
|                       | Patient 4 | 46 | 7.7 | 4 | 920 | 40 |
|                       | Patient 5 | 10 | 6.8 | 7 | 1,580 | 20 |
|                       | Patient 6 | 7 | 6.5 | 6 | 1,330 | 90 |
|                       | Patient 7 | 6 | 5.8 | NR | 900 | 10 |
|                       | Patient 8 | 13 | 4.5 | NR | 500 | 10 |
|                       | Patient 9 | 7 | 4.3 | NR | 2,330 | 10 |
|                       | Patient 10 | 12 | 4.5 | NR | 2,720 | 10 |
| **Reference range** | | | | | 700–4,500 | 90–1,000 |

NR, not reported

<sup>a</sup>Hacein-Bey-Abina et al. 2010

<sup>b</sup>Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011
repertoire. Normal counts of NK cells were detected in patients who were treated at 1–2 months of age, however, NK cell counts diminished after 18 months of treatment. Reconstitution of humoral immunity was incomplete, with 3 patients requiring ongoing intravenous immunoglobulin (IVIG) treatment. Partial immune reconstitution was achieved in an Australian TNK+ SCID-X1 patient, treated in collaboration with the French team (Ginn et al. 2005). This may have been due to the relatively low dose of γc+CD34+ cells reinfused and a reduced selective advantage for γc-expressing T cell progenitors caused by the NK+ phenotype.

The British team reported long-term outcomes for 10 patients treated in their trial, who also lacked an HLA-identical donor (Table 1.4; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011). T cell counts were within or just below the normal range for 9 patients at 4.3–8.9 years after treatment. Similar to the French trial, CD8+ T cell counts were normal for those 9 patients, while CD4+ T cell counts were normal in 6 patients. A diverse TCR repertoire was found for all 10 patients. Humoral immunity was also low in these patients, with 6 patients requiring ongoing IVIG treatment.

1.3.2 Occurrence of serious adverse events in five patients across the two independent SCID-X1 trials

The remarkable success of the French trial was lessened in 2002 when acute lymphoblastic leukaemia (ALL) developed in 2 patients, directly linked to the integrating MFG-B2γc vector (Hacein-Bey-Abina et al. 2003). ALL was subsequently reported in a further 2 patients treated in the French trial and in 1 patient treated in the British trial (Hacein-Bey-Abina et al. 2008; Howe et al. 2008). Abnormal clonal expansions were detected in these patients at 24–68 months post-treatment (Table 1.5).
<table>
<thead>
<tr>
<th>Trial</th>
<th>Patient</th>
<th>Latency (months post-treatment)</th>
<th>Immunophenotype</th>
<th>Location of vector integration</th>
<th>Dysregulated proto-oncogenes near vector integration</th>
<th>Additional somatic mutations and dysregulated genes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>French SCID-X1</td>
<td>4</td>
<td>30</td>
<td>Monoclonal immature T cell receptor γδ T cell clone</td>
<td>Reverse orientation within LMO2 intron 1</td>
<td>LMO2 mRNA levels higher than fully differentiated T cell controls</td>
<td>Partial trisomy 6 with a t(6;13) chromosomal translocation resulting in MYB over-expression; biallelic deletion of CDKN2</td>
<td>Hacein-Bey-Abina et al. 2003</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>34</td>
<td>Three different and fully differentiated γc-expressing TCRβ1, TCRβ2, and TCRβ23 clones</td>
<td>Forward orientation 3 kb upstream of LMO2 exon 1</td>
<td>LMO2 mRNA levels higher than fully differentiated T cell controls</td>
<td>A trisomy 10, and a unique SIL-TAL1 fusion caused by a deletion; heterozygous mutations in heterodimerisation domain of NOTCH1</td>
<td>Hacein-Bey-Abina et al. 2003</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>68</td>
<td>Monoclonal immature T cell receptor αβ T cell clone: CD5^+^CD1^+^CD3^-^TCRαβ^+^CD4^-^CD8^-</td>
<td>Reverse orientation 2.4 kb upstream of CCND2</td>
<td>Over-expression of CCND2, at a similar level to CCND2-rearranged T-ALL</td>
<td>Biallelic deletion of CDKN2</td>
<td>Hacein-Bey-Abina et al. 2008</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>33</td>
<td>Polyclonal immature T cell receptor αβ T cell phenotype: TdT^+^CD7^-^CD2^-^CD5^+^CD1^-^CD10^-</td>
<td>Reverse orientation 10.6 kb downstream of LMO2 exon 1; and forward orientation within intron 4 of SPAG6 and 49.5 kb downstream of BMI1</td>
<td>Over-expression of BMI1, at a similar level to leukaemias containing CALM-AF10 translocations</td>
<td>Heterozygous mutations in heterodimerisation domain of NOTCH1</td>
<td>Hacein-Bey-Abina et al. 2008</td>
</tr>
<tr>
<td>British SCID-X1</td>
<td>8</td>
<td>24</td>
<td>Monoclonal TCRβ6b clone: CD3⁻ CD4⁺CD8⁻CD7⁺CD10⁻ TdT⁻</td>
<td>Vector integration in reverse orientation, 35 kb upstream of LMO2 exon 1</td>
<td>LMO2 levels over-expressed in comparison to leukaemia panel cells, DP1 T cells and DP2 T cells</td>
<td>Mutation in NOTCH1 within heterodimerisation domain; over-expression of NOTCH1 and downstream targets HES1, c-MYC and genes within STIL-TAL locus; reduced expression of CDKN2A</td>
<td>Howe et al. 2008</td>
</tr>
<tr>
<td>CGD</td>
<td>1</td>
<td>27</td>
<td>Myelodysplasia characterised by uniformly vacuolated monocytes</td>
<td>Within MDS1-EVI1 locus</td>
<td>Over-expression of EVII, comparable to human cells with EVII rearrangement</td>
<td>Monosomy 7; genomic instability and double-strand breaks</td>
<td>Stein et al. 2010</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>42</td>
<td>Multilineage dysplasia</td>
<td>Within MDS1-EVI1 locus</td>
<td>Over-expression of EVII, comparable to human cells with EVII rearrangement</td>
<td>Monosomy 7; genomic instability</td>
<td>Stein et al. 2010</td>
</tr>
<tr>
<td>WAS</td>
<td>NR</td>
<td>41</td>
<td>T-ALL; immunophenotype not reported</td>
<td>Close to LMO2</td>
<td>NR</td>
<td>NR</td>
<td>Braun et al. 2011a</td>
</tr>
</tbody>
</table>

bp, base pairs; CGD, chronic granulomatous disease; kb, kilobases; NR, not reported; SCID-X1, X-linked severe combined immunodeficiency; T-ALL, T cell acute lymphoblastic leukaemia.

These details are taken from an American Society of Hematology abstract, since extensive analysis of the T-ALL in the WAS gene therapy trial has not yet been published.
Chapter 1: Introduction

All leukaemias displayed a T cell-like phenotype. Four of the leukaemia phenotypes were immature and 3 of the leukaemia phenotypes were monoclonal.

All patients have been treated using conventional chemotherapy for T cell acute lymphoblastic leukaemia (T-ALL). Patients 5, 7 and 10 of the French trial are in remission, with over 5 years of complete remission for patient 5 after the initial detection of leukaemia (Hacein-Bey-Abina et al. 2008). Patient 4 died 60 months after gene therapy treatment. Patient 8 of the British trial is in clinical and molecular remission (Howe et al. 2008). Both the French and British trials have been placed on voluntary hold. An international phase I/II trial of gene therapy for SCID-X1 was opened in 2010 using a redesigned vector, with centres based in London, Paris and the United States (ClinicalTrials.gov Identifiers NCT01410019, NCT01175239 and NCT01129544).

1.3.3 Vector-mediated insertional mutagenesis as the causative mechanism of leukaemogenesis

In all 5 patients who developed leukaemia in the French and British SCID-X1 trials, the development of leukaemia was directly linked to the MoMLV-based gammaretroviral vector, which integrated near proto-oncogenes and led to dysregulation of those genes. Insertional mutagenesis events may cause proto-oncogene activation by dysregulating control elements or altering a gene-product. The unique 3 (U3) region of the LTRs in the MFG-B2γc vector contains promoter-enhancer elements, and the enhancer activity of these elements can trans-activate upstream or downstream control elements in forward or reverse orientation. Insertional mutagenesis may theoretically also result from transcriptional read-through caused by a weak polyadenylation signal, disruption of silencers or mRNA stability sequences, or the production of a fusion protein with
novel characteristics via interaction of retroviral splice donor (SD) sites with splice acceptor (SA) sites in downstream genes (Baum et al. 2003; Kohn, Sadelain & Glorioso 2003). Alternatively, vector insertion within a tumour-suppressor gene may render that gene inoperative.

Strikingly, the proto-oncogene *LMO2* was implicated in 4 out of 5 of these insertional mutagenesis events (Table 1.5). The leukemic clones from Patients 4 and 5 from the French trial and Patient 8 from the British trial contained single-vector integrations near *LMO2*, leading to *LMO2* over-expression. In Patient 7 of the French trial, *CCND2* was insertionally activated by the vector. Two vector integrations were identified in the leukemic clone of Patient 10 of the French trial, 1 near *LMO2* and another near *BMI1*, causing over-expression of both genes.

A multistep mechanism of leukaemogenesis has been posited as the cause of the SAEs in both SCID-X1 trials. Whilst an activated proto-oncogene may induce signals for proliferation, such signals may be short-term and followed by senescence or induction of apoptosis. A clone containing an insertionally-activated proto-oncogene may require additional genomic lesions, in order to engraft and induce clonal proliferation (Hacein-Bey-Abina et al. 2008; Howe et al. 2008). Additional genomic lesions, including point mutations, indels and amplifications, were identified in the leukemic clones of patients in the French and British trials, which led to over-expression of proto-oncogenes such as *MYB*, *CDKN2*, *NOTCH1* and a *STIL-TAL1* fusion (Table 1.5).
1.3.4 Role of LMO2 in lymphomagenesis

The LMO2 locus encodes the LIM domain only 2 (LMO2) protein, which is involved in protein-protein interactions in transcription-factor complexes (McCormack & Rabbitts 2004). LMO2 is expressed in HPCs and has a role in early haematopoiesis, but is later down-regulated in all haematopoietic lineages except erythrocytes (Warren et al. 1994). LMO2 is a proto-oncogene associated with T-ALL, although with a low frequency and a long latency.

LMO2 has been demonstrated as a hot spot of gammaretroviral integration in human CD34+ cells (Cattoglio et al. 2007). Gammaretroviral vectors are known to integrate near actively expressed genes and LMO2 is the most prominently transcribed classical T-ALL oncogene in CD34+ cells stimulated with cytokines used in gene therapy protocols (Pike-Overzet et al. 2006; Pike-Overzet et al. 2007). Each patient in the French trial may have received up to 55 cells harbouring integrations in LMO2 (Wu et al. 2003). The failure of all of these clones to progress to leukaemia supports the multistep model of leukaemogenesis development (Section 1.3.3). Nonetheless, the over-representation of LMO2 in the SCID-X1 gene therapy adverse events points to a need for better understanding of insertional mutagenesis risks, as well as clonal dynamics and survival following gene therapy. This will be discussed in Section 1.5.

1.3.5 Possible impact of SCID-X1 background and biology on insertional mutagenesis risks

Until 2010, insertional mutagenesis had not been reported following trials of gene therapy for any other diseases, suggesting the possibility that insertional mutagenesis may have a higher prevalence in the SCID-X1 context. There may be factors relating to
the characteristics of HPCs and SCID-X1 patients that increase the risk of insertional mutagenesis. For example, SCID-X1 patients have more undifferentiated HPCs expressing LMO2, which increases the likelihood of insertions near LMO2 (Kohn, Sadelain & Glorioso 2003). Tumourogenesis is also more likely in immunocompromised patients lacking immune surveillance (Baum et al. 2004). For example, tumours that thrive in immunocompromised mice are rejected when transplanted into immunocompetent mice (Shankaran et al. 2001).

One study has suggested a possible oncogenic role for IL2Rγ. Five out of 15 mice treated with an IL2Rγ-encoding lentiviral vector developed lymphoma, in contrast to none of 15 mice treated with a green fluorescent protein (GFP)-encoding lentiviral vector (Woods et al. 2006). However, in the absence of any molecular analyses of the lymphomas, including identification of the integration sites and analyses of gene dysregulation, insertional mutagenesis cannot be excluded as a mechanism and it cannot be concluded that γc over-expression has oncogenic effects (Thrasher et al. 2006). Since then, it has been shown that γc over-expression alone is not sufficient to induce leukaemia in primary and secondary mouse recipients of wild type cells transduced with IL2Rγ-encoding vectors, although insertional activation of EVII or PRDM16 with IL2Rγ-encoding vectors can lead to murine leukaemia (Modlich et al. 2008). Perhaps the strongest evidence against IL2Rγ oncogenicity is the absence of γc over-expression in any of the leukemic clones from the SCID-X1 trials, based on analyses of γc expression levels and JAK3 activation in the patients’ blast cells (Hacein-Bey-Abina et al. 2003; Hacein-Bey-Abina et al. 2008; Howe et al. 2008).

Instead, it is more likely that the additional growth advantage of γc expression may have provided further selection for clones harbouring integrations near LMO2.
Chapter 1: Introduction

LMO2 and γc may have had a synergistic effect, with over-expression of LMO2 blocking further T cell differentiation and γc expression encouraging proliferation, predisposing such clones towards leukaemogenesis (Hacein-Bey-Abina et al. 2003). Observation of leukaemias in 2 mice containing retroviral integrations at both LMO2 and IL2Rγ may support this possibility, given the very low probability of both events occurring randomly (Davè et al. 2009; Davè et al. 2004).

1.3.6 Insertional mutagenesis in other gene therapy trials

In recent years, insertional mutagenesis has been reported in trials of gene therapy for chronic granulomatous disease (CGD) and Wiskott-Aldrich syndrome (WAS), indicating that genotoxicity is not isolated to gene therapy for SCID-X1 (Stein et al. 2010; Braun et al. 2011). Clones harbouring integrations in the EVII-MDS1 locus in both of the patients treated in the CGD trial began to dominate at 27 and 42 months post-treatment (Table 1.5). This was associated with over-expression of EVII-MDS1 and led to myelodysplasia in 1 patient and multilineage dysplasia in the other. A case of T-ALL with a vector integration near LMO2 has been reported in the German trial of gene therapy for WAS, however, mechanistic analysis has not yet been published (Braun et al. 2011). Interestingly, in a trial of gene therapy for β-thalassemia, insertional mutagenesis was associated with the survival and dominance of 1 clone, in which HMGA2 was transcriptionally activated, and which accounted for most of the therapeutic benefit and did not lead to malignancy (Cavazzana-Calvo et al. 2010).

Notably, no incidences of clonal lymphoproliferation have been reported for any of the 16 patients treated in the trials of gene therapy for ADA-SCID, even though some patients have been monitored for over 10 years. If the current outcomes of the 2
separate ADA-SCID and the 2 separate SCID-X1 trials are combined, any difference between the ADA-SCID and SCID-X1 trials in the incidence of leukaemia falls short of statistical significance (p = 0.053, Fisher’s exact test). However, as a number of these patients have been monitored for less than 5 years, it is too soon to draw any conclusions about whether patients with ADA-SCID are at less risk than those with SCID-X1.

1.4 LIMITATIONS OF PRECLINICAL MODELS FOR ASSESSING GENOTOXICITY

The occurrence of vector-mediated insertional mutagenesis in 4 independent gene therapy trials highlights the need for a better understanding of this risk, in order to minimise insertional mutagenesis in future gene therapy applications. Insertional mutagenesis is inherently associated with the use of integrating viral vectors based on retroviruses in human gene therapy. Since transduced HPCs undergo many rounds of cell division during lymphocyte ontogeny, stable integration of $IL2R\gamma$-encoding vectors is required if viral vector systems are to be used (Section 1.2.4).

Although insertional mutagenesis had been identified as a theoretical risk, it was nonetheless considered a low risk, because oncogenesis generally requires more than 1 mutation. Insertional mutagenesis was also underestimated, in large part due to the inadequacies of small animal models to measure insertional mutagenesis and predict its frequency in humans. These limitations have prompted the development of new preclinical models and assay systems to quantify the risk of insertional mutagenesis and to measure possible safety improvements in re-designed vectors. However, even with a well-designed study conducted using the best available preclinical models, a degree of
translational uncertainty is inescapable for complex risks such as insertional mutagenesis.

1.4.1 Assessments of insertional mutagenesis prior to serious adverse events in the SCID-X1 trials

Prior to the cases of leukaemia in the SCID-X1 trials, the frequency of activating an oncogene on a single allele following vector integration had been estimated as $2 \times 10^{-7}$ from in vitro testing of growth-factor independence, using TF-1 cells that require granulocyte-macrophage colony stimulating factor (GM-CSF) for growth (Stocking et al. 1993). The French team had tested their vector and protocol pre-clinically using cell lines and primary bone marrow isolated from SCID-X1 patients (Hacein-Bey et al. 1996; Hacein-Bey et al. 1998; Hacein-Bey et al. 2001). Estimates that the risk of insertional mutagenesis was extremely low were supported by animal studies and clinical trial data.

Following the report of the first 2 SAEs in the French trial, the American Society of Gene and Cell Therapy (ASGCT) appointed an Ad Hoc Subcommittee to comprehensively review existing literature on the use of retroviral vectors in HPCs in murine models, large animal models and clinical trials. The Subcommittee noted that follow-up of animals exceeded 6 months in only 30% of studies (Kohn, Sadelain, Dunbar, et al. 2003). Although no tumours were reported in the 5 gene transfer studies using γc-deficient mice, differences between the mouse model and human clinical trials limited the utility of those studies for predicting toxicity. Even though 1 possible case of murine insertional mutagenesis had been identified across all studies that included 4,846 mice, insufficient follow-up meant that it was impossible to infer a frequency of insertional mutagenesis. Since well-designed prospective studies constitute the best
means of evaluating the risk of insertional mutagenesis, the Subcommittee recommended that future studies should maintain treated mice for at least 12 months. The classic 4 month follow-up was viewed as insufficient for assessing genotoxicity-related oncogenicity.

The first report of murine leukaemia caused by retroviral vector integration was published in 2002, and only months before the French team reported the first 2 cases of leukaemia (Li et al. 2002). In the abnormal clone, a single vector integrated near the murine transcription factor $EVII$ resulting in over-expression. $EVII$ may have functioned in tandem with the transgene, since over-expression of $EVII$ alone can immortalise immature myeloid cells but is not sufficient to cause leukaemia in the absence of co-operative mutations (Nucifora 2001). This suggested the possibility of $EVII$ functioning in tandem with the transgene, a mechanism that is strikingly similar to the hypothesised co-operation between $LMO2$ and $\gamma c$ (Section 1.3.5; Hacein-Bey-Abina et al. 2003; Davé et al. 2004; Davé et al. 2009).

Large animal models have comparable life expectancies and haematopoietic demands on HPCs to humans, but are also costly and logistically difficult to maintain. When the first SAEs from the French SCID-X1 trial were reported, no large animal studies of $IL2R\gamma$ gene transfer had been reported and there was no evidence of insertional mutagenesis in any reported retroviral gene transfer studies (Kohn, Sadelain, Dunbar, et al. 2003). Since then, $IL2R\gamma$ gammaretroviral gene transfer has been used to treat a canine SCID-X1 phenotype, although engraftment was short-term and occurred in 8 of 12 dogs treated (Kennedy et al. 2011). A thymic lymphoma was found in 1 dog, containing a single vector integration near $C20orf74$, a substrate for the apoptosis-blocking $Akt$. It is unknown, however, whether this gene was over-expressed. One case
of myeloid sarcoma has been reported in a rhesus macaque, in which tumour cells contained an integration in the anti-apoptotic \textit{BCL2-A1} (Seggewiss et al. 2006).

Whilst retroviral modification of HPCs had been used in 40 clinical trials involving 232 subjects prior to the first-reported SCID-X1 SAEs, none of those studies achieved the high levels of gene transfer nor the long-term persistence of transduced cells achieved in the SCID-X1 trials (Kohn, Sadelain, Dunbar, \textit{et al}. 2003). Furthermore, most of those subjects were cancer patients with advanced disease, which limited the potential for long-term follow-up. As the first gene therapy trial to report a high degree of therapeutic efficacy that was sustained long-term, it is likely that the French SCID-X1 trial provided the first opportunity for insertional mutagenesis to present.

1.4.2 Limitations of preclinical models

A murine model of SCID-X1 gene therapy inevitably differs from a human protocol for SCID-X1 gene therapy. For example, there can be differences with respect to transduction and transplantation protocols reflecting different biology across species. Such biological and protocol differences may include the phenotype of mice versus humans, the age of donor and recipient mice, and cytokine treatment and conditioning regimes. In contrast to the classic T\textsuperscript{NK}\textsuperscript{B\textsuperscript{+}} phenotype of SCID-X1 patients, \textit{IL2R\gamma\textsuperscript{-/+}} mice have low numbers of T cells and are completely deficient in B cells and NK cells (Cao \textit{et al}. 1995). \textit{IL2R\gamma\textsuperscript{-/-}RAG2\textsuperscript{-/-}} or \textit{IL2R\gamma\textsuperscript{-/-}RAG2\textsuperscript{-/-}C5\textsuperscript{-/-}} mice, used as recipients due to the T\textsuperscript{+} phenotype of \textit{IL2R\gamma\textsuperscript{-/-}} mice, are completely deficient in T cells, B cells and NK cells. Different markers are typically used for isolating human and murine HPCs, and different cytokines may be used for stimulating and culturing HPCs. Treatment of mice
typically involves transplanting young adult recipient mice with gene-modified bone marrow following ablation of recipient marrow via sub-lethal irradiation doses, whereas in the human trials gene-modified cells are infused into non-ablated infant recipients (Kohn, Sadelain, Dunbar, et al. 2003). Retroviral transduction efficiency in murine haematopoietic cells can be over 10-fold higher than in human trials or large animal models, and a lower cell dose is typically used in mouse studies (Baum et al. 2004). Such differences may have limited the capacity of preclinical murine studies to predict the frequency of insertional mutagenesis in human trials.

Furthermore, small animal models are generally considered to be poor predictors of long-term safety outcomes in humans. Mouse oncogenesis has been acknowledged as a poor model of human oncogenesis since spontaneous leukaemia occurs sporadically in older animals lacking genetic predisposition, daily turnover of haematopoietic cells is higher, and smaller numbers of murine HPCs and a shorter life-span could underestimate risk (Cavazzana-Calvo et al. 2005; Baum et al. 2003). Furthermore, mouse cells undergo spontaneous transformation far more readily than human cells (Boehm et al. 2005; Rangarajan & Weinberg 2003; Hahn & Weinberg 2002). Background tumour formation in mouse models may confound preclinical safety testing. For example, in 1 study that involved treating wild-type mice with a retroviral vector encoding MGMT, 6 mice in the test group of 40 mice developed a T cell lymphoma compared to zero lymphomas in a control group of 10 mice (Will et al. 2007). However, no vector copies were detected in 5 of the tumours, indicating these tumours arose independently of insertional mutagenesis.

Given that large animals such as the canine SCID-X1 model are expensive to maintain and studies involving large animals involve protracted timelines, γc-deficient
mice remain the most practical preclinical models for testing efficacy and safety. Small and large animal models have yielded valuable data for evaluating the safety and efficacy of integrating retroviral vectors, yet ultimately the most informative data will come from further clinical trials. In this context, the data generated by ongoing clinical trials using gammaretroviral vectors, particularly the ADA-SCID trials, have significant value. Expression profiling analysis of T cells from patients treated in the Italian ADA-SCID trial has not uncovered evidence of gammaretroviral-mediated gene dysregulation at 10 to 30 months after treatment (Cassani et al. 2009). Fewer than 6% of genes in 18.6% of T cell clones were dysregulated modestly at 2.8- to 5.2-fold. To date, insertional mutagenesis has not been reported for any of the patients treated in the Italian and British ADA-SCID trials, even though follow-up of many of these patients exceeds 5 years.

1.4.3 Development of improved preclinical models

The limitations of SCID-X1 mouse models for measuring genotoxicity risks and the long latency periods for tumour development have pointed to a need for improved assay systems for quantifying insertional mutagenesis. A number of in vitro and in vivo systems have consequently been developed for the purpose of testing whether improvements to vector design may reduce genotoxicity. A plasmid-based system has been developed to mimic the mechanism of insertional mutagenesis-mediated oncogene activation, in which the luciferase reporter gene, \textit{Luc}, is located downstream of a proviral vector sequence followed by an internal ribosyme entry site (Hendrie et al. 2008). This system enables comparative measurements of adjacent \textit{Luc} activation by different proviral vector constructs, based on backbones such as the murine leukaemia virus (MLV), human immunodeficiency virus type-1 (HIV-1) and foamy virus (FV)
backbones, and containing different promoter-enhancer elements, including the LTR promoter-enhancer and the cellular elongation factor-1-α (EF1α) and phosphoglycerate kinase (PGK) promoter-enhancer elements.

A second cell culture assay for insertional mutagenesis is based on the serial replating of transduced primary murine bone marrow, which produces a high frequency of immortalised cell populations, many of which contain integrations near and over-expression of PRDM1/EVI1 (Modlich et al. 2006; Modlich et al. 2009). This assay has been used to compare different vector configurations. A third in vitro assay has used a recombinant adeno-associated virus (AAV) vector in Jurkat cells to insert an expression cassette with a single retroviral LTR promoter-enhancer, flanked by LoxP sites, within the first intron of LMO2 (Ryu et al. 2008). This mimics the leukaemogenic event in patient 4 of the French SCID-X1 trial. The use of Cre-mediated exchange to substitute different expression cassettes has been used to examine the effects of insulators and compare the potential of different cellular promoter-enhancer elements to dysregulate LMO2.

Two tumour-prone mouse models have been described as assays for insertional mutagenesis. The addition of a tumour-prone phenotype to IL2Rγ−/− mice has the effect of increasing the frequency of tumourigenesis and decreasing latency, thus facilitating comparisons between different vector configurations. In 1 of these models, IL2Rγ−/− mice were crossed with tumour-prone mice deficient in Arf, a tumour suppressor gene commonly mutated in T cell leukaemia. In mice treated with a γc-encoding murine stem cell virus (MSCV)-vector encoding IL2Rγ-IRES-EGFP, there was a high incidence of leukaemia following treatment, compared to mice treated with an MSCV-EGFP vector (Shou et al. 2006). Integration sites were recovered near or within known
proto-oncogenes from T cell lymphomas, indicating insertional mutagenesis caused leukaemogenesis. This was used to suggest the \( IL2R\gamma \) background is a risk factor in SCID-X1 gene therapy in that model. A second \textit{in vivo} insertional assay involving tumour-prone mice uses \( CDKN2A^{-/-} \) mice and measures accelerated rates of myeloid tumourigenesis in wild-type mice transplanted with vector-transduced \( CDKN2A^{-/-} \) HPCs (Montini et al. 2006; Montini et al. 2009). This system has been used to benchmark the genotoxicity of redesigned gene transfer vectors based on different viral backbones (Section 1.6.4).

\textbf{1.4.4 Development of knowledge about insertional mutagenesis risks and the interpretation of read-outs from preclinical models}

Knowledge about the risk of insertional mutagenesis is difficult to obtain using the standard model for SCID-X1, \( IL2R\gamma^{-/-} \) mice. The extent to which differences between human SCID-X1 gene therapy and the mouse model impact on insertional mutagenesis read-outs is unknown. Furthermore, the infrequency of tumour development and long periods of latency required render comparisons between different vector configurations difficult. While \textit{in vitro} and \textit{in vivo} preclinical models have been developed to address this challenge, these assays do not purport to provide a direct quantification of the risks of insertional mutagenesis occurring in a human subject. Human SCID-X1 patients do not have the tumour-prone genotypes used for the \textit{in vivo} models, and the \textit{in vitro} models are removed from the complex haematological systems in which tumours arise, which may include additional mutagenic stresses in a human. Nonetheless, such assays will assist researchers, institutional ethics committees (IECs), and regulatory authorities with the ultimate goal of defining an acceptable level of risk for further trials to measure genotoxicity risks in humans.
Understanding and managing risks, such as genotoxicity, requires knowledge about the type, extent, and probability of that risk. Although empirical studies help elucidate risk knowledge, the assessment of these risks is not simply a consequence of the probability of these risks, but also a judgment about their significance and meaning. Inevitably, such assessments are a question of ethics and values. Decisions about risk consequently draw upon both subjective assessments of risk and probability statistics. Accordingly, judgments about risk, including the definition and management of acceptable risk levels in human clinical trials, are normative rather than scientific. This will be discussed further in Section 1.7.

1.5 ANALYSIS OF THE INTEGRATION BEHAVIOUR OF RETROVIRAL VECTORS

Specific proto-oncogenes, such as LMO2 and MDS1-EVI1, are over-represented in the insertional mutagenesis events that have occurred in human clinical trials (Table 1.5). An understanding of the integration behaviour of retroviral vectors and how this may lead to clonal dominance has therefore been a critical aspect of research into insertional mutagenesis. Such knowledge has also informed the design of improved vectors for gene transfer. In parallel to this research, methodologies for analysing vector integration behaviour have been developed and improved upon, with such improvements including the adoption of recent advances in NGS technology. By providing insights into the kinetics of clonal engraftment, selection and survival in both preclinical models and humans, these methodologies have been indispensible as tools for measuring the efficacy of gene therapy as well as monitoring for long-term safety.
1.5.1 Integration of gammaretroviral vectors is non-random

Although retroviral integration was historically understood to be random, a growing line of evidence has indicated that retroviruses integrate more frequently within certain chromosomal sites than others. Target specificity during retroviral integration is affected by the central core domain of retroviral integrase, and by factors such as methylation state, nuclear scaffold attachment, nucleosome structure, transcriptional activity, chromatin packaging and GC content (Laufs et al. 2003). Consequently, retroviral vectors are more likely to favour certain sites depending on the pattern of gene expression in target cells. Non-random retroviral integration preferences affected the frequency of insertional mutagenesis in the SCID-X1 gene therapy trials. Therefore, it is imperative to understand which sites may be favoured in the target cells of interest during retroviral integration in order to better understand insertional mutagenesis.

Integrating vectors based on retroviruses, such as MLV and HIV-1, are more likely to integrate in or near actively transcribed genes where chromatin is less tightly packaged and DNA is more accessible (Table 1.6; Schröder et al. 2002; Wu et al. 2003; Wagner et al. 2005). Gammaretroviral vectors also display a preference for integrating at fragile sites that correlate with chromosomal breakpoints (Bester et al. 2006). MLV has been shown to target genes highly expressed in hematopoietic progenitors and stem cells, particularly genes associated with cell-cycle control, apoptosis signalling and transcriptional regulation (Beard et al. 2007; Kustikova et al. 2007). MoMLV-based gammaretroviral vectors and not HIV-1-based lentiviral vectors target cancer-associated and cell cycle control-associated genes at the pre-transplantation stage (Montini et al. 2006).
Table 1.6: Summary of key findings of studies investigating retroviral integration behaviour and retroviral-induced clonal dominance

<table>
<thead>
<tr>
<th>Finding and significance</th>
<th>Total integration sites recovered, integration site methodology, cell type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Retroviral vector integration is not randomly distributed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV-1 preferentially integrates into genes, especially genes expressed following HIV infection</td>
<td>524, LM-PCR, human T cell line SupT1</td>
<td>Schröder et al. 2002</td>
</tr>
<tr>
<td>Integration of a retroviral vector based on the Friend mink cell focus-forming/murine embryonic stem cell virus is not randomly distributed across all human chromosomes, with over-representation of chromosomes 6, 17 and 19 and under-representation of chromosome 4. Specific chromosomal regions are preferentially targeted</td>
<td>156, LM-PCR, human CD34⁺ peripheral blood progenitor cells transplanted into NOD/SCID mice</td>
<td>Laufs et al. 2003</td>
</tr>
<tr>
<td>Retroviruses such as MLV target actively transcribed genomic regions packaged into euchromatin</td>
<td>1,282, LM-PCR, HeLa cells</td>
<td>Wu et al. 2003</td>
</tr>
<tr>
<td>Retroviral vectors are more likely to integrate near CpG islands, commonly associated with transcriptional start sites: 58% of HIV-1 integrations and 34% of MLV integrations occur in actively transcribed genes, compared to 22.4% of random integrations</td>
<td>1,282, LM-PCR, HeLa cells</td>
<td>Wu et al. 2003</td>
</tr>
<tr>
<td>Preference of retroviral vector based on the Friend mink cell focus-forming/murine embryonic stem cell virus to integrate near the transcriptional start regions, within CpG islands, within Alu repeat regions and within the first intron of genes</td>
<td>186, LM-PCR, human CD34⁺ peripheral blood progenitor cells transplanted into NOD/SCID mice</td>
<td>Laufs et al. 2004</td>
</tr>
<tr>
<td>Retroviral vector based on the Friend mink cell focus-forming/murine embryonic stem cell virus is more likely to integrate near actively transcribed genes</td>
<td>117, LM-PCR, human CD34⁺ peripheral blood progenitor cells</td>
<td>Wagner et al. 2005</td>
</tr>
</tbody>
</table>
68% of HIV-1-based lentiviral vector integrations within RefSeq genes, mostly within introns

Analysis of 380 MLV-based gammaretroviral and 235 HIV-based lentiviral vector integrations isolated from the peripheral blood of 8 baboons and 2 pigtailed macaques showed the preference of both vectors for integration within genes, particularly proto-oncogenes, genes involved in growth and survival, and genes highly expressed in haematopoietic cell populations, with MLV-vectors targeting the transcriptional start site. No observation of a hotspot at MDS1/EVI1.

MLV-based gammaretroviral vectors and not HIV-1-based lentiviral vectors appear to target cell-type specific transcription binding via a mechanism involving MLV integrase and the LTR enhancer, possibly through the tethering of the gammaretroviral preintegration complex by transcription factors. Overrepresentation of these integrations is eliminated in gammaretroviral vectors containing U3 deletions.

Preferential integration of the MoMLV-based gammaretroviral vectors within genes and near the transcriptional start site demonstrated using samples taken from patients treated in French and British SCID-X1 trials. Analysis revealed clustering of integrations at common integration sites, representing a quarter of integrations isolated from patients in the French trial, and suggesting integrations at these loci can influence the engraftment, survival and proliferation of clones. Integrations from post-transplantation samples displayed a clear preference for growth-promoting genes compared to pre-transplantation samples. The most commonly represented sites recovered from patients treated in the French trial were CCND2, ZNF217 and LMO2, which represented 3% of all clones but only \(7 \times 10^{-7}\) of the genome.

| 68% of HIV-1-based lentiviral vector integrations within RefSeq genes, mostly within introns | 57, LM-PCR, human CD34\(^+\) peripheral blood progenitor cells transplanted into NOD/SCID mice | Laufs et al. 2006 |
| Analysis of 380 MLV-based gammaretroviral and 235 HIV-based lentiviral vector integrations isolated from the peripheral blood of 8 baboons and 2 pigtailed macaques showed the preference of both vectors for integration within genes, particularly proto-oncogenes, genes involved in growth and survival, and genes highly expressed in haematopoietic cell populations, with MLV-vectors targeting the transcriptional start site. No observation of a hotspot at MDS1/EVI1. | 665, LAM-PCR, NHP CD34\(^+\) transplanted into baboons and pigtailed macaques | Beard et al. 2007 |
| MLV-based gammaretroviral vectors and not HIV-1-based lentiviral vectors appear to target cell-type specific transcription binding via a mechanism involving MLV integrase and the LTR enhancer, possibly through the tethering of the gammaretroviral preintegration complex by transcription factors. Overrepresentation of these integrations is eliminated in gammaretroviral vectors containing U3 deletions | 795, LM-PCR, human CD34\(^+\) cord blood progenitor cells | Felice et al. 2009 |
| Preferential integration of the MoMLV-based gammaretroviral vectors within genes and near the transcriptional start site demonstrated using samples taken from patients treated in French and British SCID-X1 trials. Analysis revealed clustering of integrations at common integration sites, representing a quarter of integrations isolated from patients in the French trial, and suggesting integrations at these loci can influence the engraftment, survival and proliferation of clones. Integrations from post-transplantation samples displayed a clear preference for growth-promoting genes compared to pre-transplantation samples. The most commonly represented sites recovered from patients treated in the French trial were CCND2, ZNF217 and LMO2, which represented 3% of all clones but only \(7 \times 10^{-7}\) of the genome | 572 from French trial and 439 from British SCID-X1 trials, LAM-PCR and 454 sequencing, human CD34\(^+\) transplanted autologously into SCID-X1 patients | Deichmann et al. 2007; Schwarzwaelder et al. 2007 |
Integration by HIV-1-based lentiviral vectors show a strong preference for and MLV-based gammaretroviral vectors show a weaker preference for integration within transcription units.

<table>
<thead>
<tr>
<th>Integration</th>
<th>Methods</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-1-based</td>
<td>160,232, LM-PCR with 454 sequencing, murine Sca1(^+) progenitor cells transplanted into mice</td>
<td>Wang et al. 2008</td>
</tr>
<tr>
<td>MLV-based</td>
<td>Over 32,631 MLV integrations and 28,382 HIV-1 integrations, LM-PCR and 454 sequencing human CD34(^+) umbilical cord progenitor cells</td>
<td>Cattoglio et al. 2010</td>
</tr>
</tbody>
</table>

2. **Hotspots of retroviral integration**

A search of over 3000 retroviral integrations that induced leukaemia in mice, listed in the Mouse Retroviral Tagged Cancer Gene Database, identified 2 leukaemias with integrations at \(LMO2\) and 2 with integrations at \(IL2R\gamma\), 1 of which contained integrations at both loci. Since the likelihood of randomly identifying a leukaemia with integrations at both \(LMO2\) and \(IL2R\gamma\) is \(10^{-6}\), this suggests co-operativity between \(LMO2\) and \(IL2R\gamma\) for leukaemogenesis.

<table>
<thead>
<tr>
<th>Hotspots</th>
<th>Methods</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>(LMO2) and (IL2R\gamma)</td>
<td>Not applicable(^a)</td>
<td>Davè et al. 2004</td>
</tr>
<tr>
<td>14 independent intragenic MLV retroviral integrations identified within introns 1 and 2 of the (MDS1/EVI1) complex in 9 out of 22 rhesus macaques, suggesting (MDS1/EVI1) may be a retroviral integration hot-spot or integration at this site may affect haematopoietic potential or survival</td>
<td>702, inverse PCR and LAM-PCR, NHP CD34(^+) transplanted into rhesus macaques</td>
<td>Calmels et al. 2005</td>
</tr>
<tr>
<td>MLV-based gammaretroviral vectors are more likely to integrate in common fragile sites, which correlate with chromosomal breakpoints. (LMO2) is within a common fragile site</td>
<td>1,486, LAM-PCR, human CD34(^+) pre-transplant, peripheral CD3(^+) post-transplant cells from French SCID-X1 trial and HeLa cells</td>
<td>Bester et al. 2006</td>
</tr>
<tr>
<td>Gammaretroviral vectors integrate near genes regulating cell growth and proliferation, show a high frequency of integration hot spots and a propensity to integrate at recurrent sites. Proto-oncogenes and cancer-associated common integration sites are hot spots of gammaretroviral integration</td>
<td>1,030, LM-PCR and LAM-PCR, human CD34(^+) cells</td>
<td>Cattoglio et al. 2007</td>
</tr>
</tbody>
</table>
HIV-1-based lentiviral vector integrations occur preferentially within AT rich regions of the genome that are flanked by GC rich regions

HIV-1-based lentiviral vector integrations are strongly enriched within active genes and in outward facing DNA major grooves, but not near proto-oncogenes or within tumour-suppressor genes

Prolonged ex vivo expansion of NHP CD34+ progenitors increases the frequency of MLV retroviral integrations in MDS1/EVI1, compared to previously reported data from Calmels et al. (2005)

Over 3,500 hotspots of MLV-based retroviral integration co-localise with regulatory elements (including transcription start sites, promoters, enhancers and evolutionarily conserved non-coding regions), within or around genes involved in haematopoietic functions, and with regions with epigenetic signatures of active transcription

The enrichment of MoMLV-based gammaretroviral vector integration near epigenetic markers associated with promoters and active transcription units was stable over time. Significant clustering of integration sites was observed, which correlated positively with clonal abundance. Clusters generally occurred at loci which either promoted cell growth and persistence or which were favoured for initial integration

3. Gammaretroviral integration can induce clonal dominance

MLV-based gammaretroviral integration can induce clonal dominance through insertion near and activation of genes involved in self-renewal or survival, such as EVII, CCND3 and genes in the HOXB cluster

MSCV-based retroviral integration near transcription factor homologs EVII and PRDM16 promotes immortalisation of bone marrow cells. EVII over-expression alone can immortalise immature myeloid cells without cooperating mutations

<table>
<thead>
<tr>
<th>HIV-1-based lentiviral vector integrations</th>
<th>Over 3,500 hotspots of MLV-based retroviral integration</th>
<th>Over 32,000 MLV integrations, LM-PCR and 454 sequencing human CD34+ umbilical cord progenitor cells</th>
<th>Prolonged ex vivo expansion of NHP CD34+ progenitors increases the frequency of MLV retroviral integrations in MDS1/EVI1, compared to previously reported data from Calmels et al. (2005)</th>
<th>Over 3,500 hotspots of MLV-based retroviral integration co-localise with regulatory elements (including transcription start sites, promoters, enhancers and evolutionarily conserved non-coding regions), within or around genes involved in haematopoietic functions, and with regions with epigenetic signatures of active transcription</th>
<th>The enrichment of MoMLV-based gammaretroviral vector integration near epigenetic markers associated with promoters and active transcription units was stable over time. Significant clustering of integration sites was observed, which correlated positively with clonal abundance. Clusters generally occurred at loci which either promoted cell growth and persistence or which were favoured for initial integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>40,569, LM-PCR with 454 sequencing, Jurkat cells</td>
<td>7,782, LM-PCR with 454 sequencing, human CD4+ T cells</td>
<td>34, LAM-PCR, NHP CD34+ transplanted into rhesus macaques</td>
<td>Over 32,000 MLV integrations, LM-PCR and 454 sequencing human CD34+ umbilical cord progenitor cells</td>
<td>Over 3,500 hotspots of MLV-based retroviral integration co-localise with regulatory elements (including transcription start sites, promoters, enhancers and evolutionarily conserved non-coding regions), within or around genes involved in haematopoietic functions, and with regions with epigenetic signatures of active transcription</td>
<td>The enrichment of MoMLV-based gammaretroviral vector integration near epigenetic markers associated with promoters and active transcription units was stable over time. Significant clustering of integration sites was observed, which correlated positively with clonal abundance. Clusters generally occurred at loci which either promoted cell growth and persistence or which were favoured for initial integration</td>
</tr>
</tbody>
</table>

| 9,767, LM-PCR with 454 sequencing, peripheral blood post-transplant cells from French SCID-X1 trial | 3. Gammaretroviral integration can induce clonal dominance | 3. Gammaretroviral integration can induce clonal dominance through insertion near and activation of genes involved in self-renewal or survival, such as EVII, CCND3 and genes in the HOXB cluster | 3. Gammaretroviral integration can induce clonal dominance through insertion near and activation of genes involved in self-renewal or survival, such as EVII, CCND3 and genes in the HOXB cluster | 3. Gammaretroviral integration can induce clonal dominance through insertion near and activation of genes involved in self-renewal or survival, such as EVII, CCND3 and genes in the HOXB cluster | 3. Gammaretroviral integration can induce clonal dominance through insertion near and activation of genes involved in self-renewal or survival, such as EVII, CCND3 and genes in the HOXB cluster |

Cooperation between a SOX4 retroviral transgene and proto-oncogenes near the vector integration led to myeloid leukaemia in 13 transplanted mice

A rhesus macaque transplanted with CD34⁺ cells transduced with an eGFP-expressing MSCV-based retroviral vector developed acute myeloid leukaemia 5 years post-transplantation; leukemic clone contained integrations in and over-expression of the anti-apototic gene BCL2-A1

Retroviral integrations resulting in benign or malignant clonal dominance in long-term murine studies are located in or near proto-oncogenes (22.5%), signalling genes (49.6%) and genes of unknown function (27.9%)

MLV-based gammaretroviral integrations near the 5' end of cancer-associated genes were enriched in mice that subsequently developed lymphoproliferation, compared to gammaretroviral integrations in murine embryonic fibroblasts

Clonal dominance was observed in recipients of gammaretrovirally-transduced LinSca1⁺c-Kit⁺ HPCs but not in recipients of retrovirally-transduced LinSca1⁺c-Kit⁺ HPCs with less intrinsic repopulation potential

Non-malignant clonal expansions identified in the CCND2 promoter in 2 patients and in the third intron of HMGA2 in 2 patients treated in the SCID-X1 trial

| 4. Integration pattern of gammaretroviruses compared to lentiviruses |
|---|---|---|
| 16.8% of MLV integrations occur in the 1 kb region surrounding 27 704 known human CpG islands associated with vertebrate transcriptional starts, compared to 2.1% of HIV-1 integrations and random integrations | 1,282, LM-PCR, HeLa cells | Wu et al. 2003 |
| MLV-based vectors inserted near the transcriptional start, while SIV-based vectors inserted throughout the whole of transcription units and in gene-dense regions | 992, LAM-PCR, NHP CD34⁺ progenitor cells transplanted into rhesus macaques | Hematti et al. 2004 |

<table>
<thead>
<tr>
<th>Integration Type</th>
<th>Linkage to Gene Expression</th>
<th>Studies</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-1-based lentiviral</td>
<td>Predominantly in gene rich regions of chromosomes</td>
<td>40</td>
<td>[1] H-1-based lentiviral integrations were predominantly in gene rich regions of chromosomes, and particularly in actively transcribed genes. MLV-based gammaretroviral vectors showed a strong preference to integrate around transcriptional start sites. Avian sarcoma leukosis virus-based vectors showed a weak preference for actively transcribed genes and no preference for transcriptional start sites. 20% of HIV-1-based lentiviral integrations were within 10 kb of transcriptional start sites, a significantly lower proportion than gammaretroviral integrations. Analysis of 1,030 MLV-based gammaretroviral vector integrations and 849 HIV-1-based lentiviral vector integrations indicated gammaretroviral vectors are more likely to integrate near transcriptional start sites. Gammaretroviral vectors are more likely to target genes regulating cell growth and proliferation, and show a high frequency of integration hot spots or propensity to integrate at recurrent sites. Proto-oncogenes and cancer-associated common integration sites are hot spots of gammaretroviral and not lentiviral integration. MLV-based gammaretroviral vectors preferentially integrate near transcriptional start sites and near CpG islands, whereas integration of HIV-1-based lentiviral vectors is disfavoured near CpG islands. Sustained and highly polyclonal reconstitution in 3 rhesus macaques followed for 4 years after treatment with a SIV lentiviral vector, with no integrations in MDS1/EVI1, PRDM16 or EVII, in contrast to previous report by Calmels et al. (2005) of NHPs treated with MLV-based gammaretroviral vector.</td>
</tr>
<tr>
<td>Gammaretroviral</td>
<td>Around transcriptional start sites</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>MLV-based gammaretroviral</td>
<td>Near transcriptional start sites and near CpG islands</td>
<td>160,232, LM-PCR with 454 sequencing, murine Sca1^+ progenitor cells transplanted into mice</td>
<td>Wang et al. 2008</td>
</tr>
<tr>
<td>HIV-1-based lentiviral</td>
<td>Disfavoured near CpG islands</td>
<td>519, LAM-PCR with 454 sequencing, NHP CD34^+ transplanted into rhesus macaques</td>
<td>Kim et al. 2009</td>
</tr>
</tbody>
</table>

**References:**

1. [Meta-analysis of retroviral integrations listed in the Mouse Retroviral Tagged Cancer Gene Database.](#)
A non-random distribution of MLV-based vector integrations was strikingly demonstrated in non-human primates, in which a high frequency of integrations near \( EVII \) has been observed (Calmels et al. 2005). In 9 rhesus macaques, 14 independent integrations were found within introns 1 and 2 of the \( MDS1/EVII \) gene complex, a region involved in chromosomal translocations in human myeloid leukaemias. The probability of observing 14 random integration events in a single locus is extremely rare, at \( 1.7 \times 10^{-34} \), which is significantly lower than this observed frequency. Despite this bias towards insertion in \( MDS1/EVII \), malignant transformation due to \( MDS1/EVII \) dysregulation was not detected in any of the 9 monkeys.

More significantly, analyses of samples taken from patients treated in both the French and British trials revealed clustering of integrations at common integration sites, which represented a quarter of all vector integrations retrieved from patients in the French trial (Deichmann et al. 2007; Schwarzwaelder et al. 2007). This suggested integration at these loci can influence the engraftment, survival and proliferation of clones. Integrations from post-transplantation samples displayed a clear preference for growth-promoting genes compared to pre-transplantation samples. The most commonly represented sites were \( CCND2, ZNF217 \) and \( LMO2 \), which represented 3% of all clones but only \( 7 \times 10^{-7} \) of the genome.

Based on the evidence above that retroviral vectors do not integrate randomly, the risk of integrating within 200 kilobases (kb) of a given proto-oncogene may increase from previous estimates of \( 2 \times 10^{-7} \) to \( 10^{-3} \text{–} 10^{-2} \) (Baum et al. 2003). However, the frequency of insertion near a proto-oncogene does not necessarily equate to the frequency of consequently activating a proto-oncogene, or even the frequency of oncogenesis. These latter 2 events are context-dependent and involve complex
molecular mechanisms. The \textit{LMO2} locus is actively transcribed in HPCs and while it has been estimated that patients of SCID-X1 gene therapy may have each received up to 55 vector integrations in the 5' region of \textit{LMO2} (Wu \textit{et al.} 2003), malignant transformation involving \textit{LMO2} over-expression was observed in 5 out of 20 patients. Accordingly, it has become essential to investigate the effect of integrating retroviral vectors on clonal dynamics and survival.

\subsection*{1.5.2 Gammaretroviral integration can trigger clonal dominance}

Integration of replication-incompetent retroviruses can trigger non-malignant clonal expansion in HPCs by improving selection fitness of clones, possibly by a gain of long-term engraftment or by immortalisation. This may be through inserting near and activating genes involved in self-renewal or survival, such as \textit{EVI1}, \textit{CCND3} and genes in the \textit{HOXB} cluster (Table 1.6; Kustikova \textit{et al.} 2005). The capacity of retroviruses such as MoMLV, murine stem cell virus and simian virus 40 to trigger clonal dominance has been exploited for the purpose of identifying novel proto-oncogenes (Du, Jenkins, \textit{et al.} 2005; Du, Spence, \textit{et al.} 2005; Li \textit{et al.} 2007). The propensity to progress to clonal dominance also relates to the stem cell potential of retrovirally-marked cells. Whereas progressive clonal dominance was observed in mice transplanted with cells derived from Lin\textsuperscript{-}Sca1\textsuperscript{+}c-Kit\textsuperscript{+} HPCs, no clonal imbalance was observed in mice transplanted with Lin\textsuperscript{-}Sca1\textsuperscript{-}cKit\textsuperscript{+} HPCs which have comparatively less intrinsic repopulation potential (Kustikova \textit{et al.} 2009). Higher donor chimerism was observed in the bone marrow, peripheral blood and spleen of mice that received lentivirally transduced long-term repopulating HPCs compared to mice that received gammaretrovirally transduced long-term repopulating HPCs.
The murine \textit{EVII} locus and the human and non-human primate \textit{MDS1/EVII} loci have frequently been identified as loci that can promote clonal dominance (Table 1.6; Du \textit{et al.} 2005; Kustikova \textit{et al.} 2005; Kustikova \textit{et al.} 2007). Over-expression of murine and human \textit{EVII} is known to be involved in myelodysplasia, possibly by blocking interferon-dependent induction of the tumour-suppressor \textit{PML} (Buonamici \textit{et al.} 2003; Buonamici \textit{et al.} 2004; Buonamici \textit{et al.} 2005). Dominant and initially non-malignant clones containing activating integrations in \textit{MDS1/EVII}, \textit{PRDM16} and \textit{SETBP1} were detected in both of the patients treated in the CGD gene therapy trial (Ott \textit{et al.} 2006). Although those clones lacked self-renewal capacity when plated and replated in methylcellulose at an early time-point, they subsequently progressed to malignancy in both patients, with \textit{EVII} over-expression accompanied by genomic instability (Stein \textit{et al.} 2010).

Non-malignant expanded clones may have a subtle selective advantage that does not result in uncontrolled proliferation in the absence of an additional factor. At present, such factors are poorly understood but are likely to include the acquisition of further somatic mutations.

1.5.3 Gammaretroviral vectors preferentially integrate near transcriptional start sites compared to lentiviral vectors

In addition to the non-random nature of retroviral integration, there is a strong line of evidence showing that MLV-based gammaretroviral vectors are more likely to integrate within 5 kb upstream or downstream of the transcriptional start site of genes, where dysregulation of cellular genes is most likely (Table 1.6). Analyses of samples taken from the patients treated in both the French and British trials indicated the integration sites of the MoMLV-based gammaretroviral vectors were clustered around the
transcriptional start site (Deichmann et al. 2007; Schwarzwaelder et al. 2007). This preferential integration may stem from targeting by gammaretroviral vectors of cell-type specific transcription factor binding sites via a mechanism involving MLV integrase and the LTR promoter-enhancer, through tethering of the gammaretroviral preintegration complex by transcription factors binding to motifs in the U3 region of the LTR (Felice et al. 2009). Integrations within transcription factor binding sites occur less frequently for SIN vectors in which this U3 region is deleted.

By contrast, lentiviral vectors, such as those based on HIV-1 or simian immunodeficiency virus (SIV), also favour integration within genes but display a more random pattern of integration across the entire transcriptional unit (Table 1.6). The preference of gammaretroviral vectors and not lentiviral vectors to integrate near the transcriptional start site has been demonstrated in HeLa cells (Wu et al. 2003), human CD34+ cells (Cattoglio et al. 2007; Felice et al. 2009), primate HPCs (Hematti et al. 2004; Calmels et al. 2005; Beard et al. 2007; Kim et al. 2009), human peripheral blood progenitor cells transplanted in mouse xenografts (Laufs et al. 2004), murine wild type bone marrow (Kustikova et al. 2007), and mouse CDKN2A−/− Lin− cells transplanted in wild-type mouse recipients (Montini et al. 2006). Unlike gammaretroviral vectors, lentiviral vectors do not preferentially integrate near cancer-, cell cycle- and cell death-related genes (Montini et al. 2009).

The tendency of MLV-based gammaretroviral vectors to integrate near the transcriptional start site where vector integration is most likely to affect gene regulation suggests gammaretroviral vectors are theoretically more likely than HIV-1-based lentiviral vectors to cause insertional mutagenesis. The ‘safer’ integration profile of HIV-1-based lentiviral vectors may offer a safety advantage over the MoMLV-based
gammaretroviral vectors used in the SCID-X1 trials, and is a rationale for selecting a lentiviral backbone. Studies comparing the relative genotoxicity of gammaretroviral vectors and lentiviral vectors will be discussed further in Section 1.6.4.

A more random integration profile with respect to intragenic integration has been demonstrated using FV-based vectors, in contrast to both gammaretroviral and lentiviral vectors (Trobridge et al. 2006). To date, however, only a handful of preclinical studies using FV vectors have been reported (Bauer et al. 2008; Bauer et al. 2011; Trobridge et al. 2012).

1.5.4 Limitations of methods for analysing retroviral integrations

Several methodologies have been described for identifying the integration sites of retroviral vectors such as gammaretroviral vectors and lentiviral vectors. Of these methods, the most frequently used are ligation-mediated PCR (LM-PCR) and linear amplification-mediated PCR (LAM-PCR). In LM-PCR, genomic DNA containing retroviral integration sites is restriction endonuclease-digested and the resultant overhangs are ligated to a linker cassette of known sequence (Mueller & Wold 1989; Schmidt et al. 2001). This enables amplification of the junction between the vector and genomic DNA using primers specific for the vector LTR and linker cassette. LAM-PCR is a modification of this method that includes an initial pre-amplification of the vector-genomic junctions using an LTR-specific primer and magnetic selection of those products prior to restriction digest and linker cassette ligation (Schmidt et al. 2002). LAM-PCR is generally considered to be more sensitive than LM-PCR, and has successfully identified integration sites at the single-cell level, estimated to be present at 0.1%–1% of a population of transduced cells (Schmidt et al. 2002).
Whilst LM-PCR and LAM-PCR have provided useful methodologies for analysing clonal diversity and both have led to insights into the safety and efficacy of haematopoietic gene therapy, there are nevertheless inherent limitations to these methods that impact upon their utility and validity. The first of these limitations relates to the reliance of both methods on restriction endonuclease digestion and multiple cycles of PCR. Since smaller targets are more efficiently PCR-amplified than larger ones, the efficiency of identifying integration sites depends on their proximity to the recognition motif for the restriction enzyme used. Integration sites located greater than 1 kb away from a recognition motif are unlikely to be amplifiable (Gabriel et al. 2009). Conversely, short vector-genomic fragments amplify more efficiently than longer ones and may be over-represented. However, the genomic sequence in short vector-genomic fragments may be too short to map uniquely to the genome. Thus, multiple restriction enzymes recognising a range of sequence motifs are employed as a means of overcoming these biases, and over 85% of the genome may be accessible to integration site analysis if 4 or 5 enzymes recognising non-GC motifs are used (Gabriel et al. 2009). Further limitations relate to mapping of the genomic portion of vector-genomic junctions to the genome. Some integration sites may be unidentifiable if they map to multiple loci, or if they are located in the unaligned portion of the human genome, which constitutes approximately 7% of the genome (Gabriel et al. 2009).

The consequence of these limitations for analyses of clonal diversity is that certain clones within the repertoire may be missed and read-outs of clonal contributions can only be semi-quantitative at best. For example, LAM-PCR has been shown to miss the leukemic clone from one of the patients treated in the French SCID-X1 trial when the restriction motif was over 1 kb from the restriction motif used, even though that clone was present at 50-fold higher abundance than other clones (Gabriel et al. 2009).
In the same study, LM-PCR retrieved only 45% of integration sites contained in a known artificial mixture. The standard LAM-PCR method failed to detect 30%–40% of clones within defined artificial mixtures, and measurements of the relative abundance of clones deviated by up to 60-fold from the known abundance (Harkey et al. 2007). A modified LAM-PCR method using additional restriction motifs increased the identifiable proportion of clones to over 90%. Another study reported that LAM-PCR detected only 55% of clones within a defined artificial mixture (Wu et al. 2013).

Although integration site analysis is crucial for research into insertional mutagenesis and analyses of the safety profile of redesigned vectors for gene transfer, the flaws of the commonly used methods inevitably limit the degree of conclusiveness of the findings of such studies. Consequently, some uncertainty persists about insertional mutagenesis and the safety of redesigned vectors.

1.5.5 Development of non-restrictive methods

As part of the gene therapy field’s efforts to understand insertional mutagenesis and develop improved methods for monitoring clonal diversity and safety, new integration site methodologies have been developed that purport to address some of the biases and limitations associated with LM-PCR and LAM-PCR. Non-restrictive LAM-PCR (nrLAM-PCR) eliminates the restriction endonuclease step of LAM-PCR by ligating a single-stranded linker cassette to the pre-amplified vector-genomic junction in single-stranded form (Gabriel et al. 2009). This method has identified a representative distribution of integration sites from a polyclonal sample. Several other restriction-free strategies have been developed, and while the retrieval of integration sights using these methods is less biased than LM-PCR and LAM-PCR, their quantitative potential
remains unproven (Pule et al. 2008; Paruzynski et al. 2010; Brady et al. 2011; Wu et al. 2013). Indeed, one such restriction-free method was non-quantitative with respect to the relative contribution of clones within a mixed sample, and was impugned by biases that were independent of the use of restriction endonucleases and may relate to polymerases (Wu et al. 2013).

1.5.6 Application of next generation sequencing technology to integration site analysis

Until recently, the sensitivity of integration site methodologies for detecting unique integration sites within a complex population was restricted to the number of final PCR products that were subcloned and Sanger-sequenced. The advent of NGS technology has opened the potential to obtain the sequence identity of thousands of integration sites from a single sample. The NGS platform that has been most frequently used for this application is 454 pyrosequencing, which can generate over 1 million reads from a single full plate run.

The increased resolution to integration site analysis that is enabled by the high data output from NGS has provided new insights into vector integration profiles and clonal dynamics (Table 1.6; Wang et al. 2007; Wang et al. 2008; Wang et al. 2009; Wang et al. 2010). The most clinically significant of these studies examined the long-term clonal dynamics of T cell samples taken from patients treated in the French SCID-X1 trial, and analysed a total of 9,767 unique integration sites (Wang et al. 2010). A correlation was identified between the dose of transduced cells received by patients and the clonal diversity of reconstituted T cells. Importantly, this longitudinal analysis revealed that the diversity of clones, clustering of integration sites and expansion of certain clonal populations were established at an early time point after gene therapy.
The application of NGS technology to integration site analysis has also been augmented by the development of high-throughput computational tools for mapping integration sites and characterising the nearby genomic features, including the identity and function of nearby genes (Peters et al. 2008; Appelt et al. 2009).

Whilst significant biological insights have been facilitated by the high throughput of NGS analysis of integration site libraries, such libraries are likely to be limited by biases introduced by the use of restriction endonucleases and multiple cycles of PCR amplification during library preparation (Section 1.5.4). Therefore, the need for methods enabling minimally biased analysis of clonal contributions and dynamics remains.

1.6 STRATEGIES FOR DESIGNING SAFER VECTORS FOR SCID-X1 GENE THERAPY

Research into the risk of insertional mutagenesis has informed strategies for developing safer vectors for use in gene therapy for SCID-X1, as well as other diseases affecting the haematopoietic compartment. These strategies have included adopting a self-inactivating (SIN) configuration, changing the vector backbone from a gammaretroviral backbone to a lentiviral backbone, and careful selection of an internal heterologous promoter-enhancer element. The effects of these modifications have been tested in preclinical models and safety assays.

1.6.1 Self-inactivating vector configuration

The strong enhancer activity of the LTR promoter-enhancer elements within the gammaretroviral vectors used in the French and British SCID-X1 trials activated proto-
oncogenes near the integration sites in the malignant clones of the patients who
developed leukaemia. The promoter-enhancer element in the retroviral LTR region has
strong constitutive enhancer activity that can influence genomic regions as large as 90
kb (Bartholomew & Ihle 1991). In a SIN vector, deletions in the U3 region of the LTRs
include the promoter and enhancer elements, and transgene expression is controlled by
an internal cellular regulatory element, allowing tighter regulation of expression (Dull
et al. 1998). SIN gammaretroviral vectors containing no internal promoter have been
shown to have a primarily neutral effect on adjacent genes and do not appear to induce
clonal dominance (Cornils et al. 2009). While one study has shown that U3 deletions
reduce the likelihood of gammaretroviral integrations within transcription factor
binding sites (Felice et al. 2009), another has shown that these deletions do not affect
the integration preferences of SIN gammaretroviral vectors (Moiani et al. 2013).

The reduced propensity of SIN vectors to cause insertional activation has been
demonstrated in a number of different systems. In the plasmid-based insertional
mutagenesis assay (Section 1.4.3), significantly reduced Luc activity was demonstrated
with SIN MLV, HIV and FV constructs containing an internal murine PGK promoter
compared to an MLV-based vector with an intact LTR promoter-enhancer (Hendrie et
al. 2008). Using the CDKN2A−/− tumour-prone mouse model, comparisons were made
between SIN gammaretroviral and SIN lentiviral vectors containing either the PGK or
the strong spleen focus-forming virus (SFFV) promoter-enhancer elements, and
gammaretroviral and lentiviral vectors containing the SFFV promoter-enhancer in place
of the LTR U3 region (Montini et al. 2009). Transcriptionally active LTRs were
identified as the major determinants of insertional mutagenesis, and SIN lentiviral and
gammaretroviral vectors with an internal PGK promoter-enhancer did not accelerate
tumourigenesis.
The international phase I/II trial of gene therapy for SCID-X1 is testing a SIN gammaretroviral vector containing the intron-less form of the EF1α (EFS) promoter as the internal promoter to control γc expression (Thornhill et al. 2008). However, given the evidence that lentiviral vectors are less likely to target the transcriptional start site of genes, SIN lentiviral vectors may be a safer alternative to SIN gammaretroviral vectors.

1.6.2 Development of HIV-1-based lentiviral vectors for gene therapy

Lentiviral vectors show promise for use in gene therapy targeting the haematopoietic compartment. Lentiviral vectors can transduce both dividing and non-dividing cells (Naldini, Blömer, Gallay, et al. 1996; Naldini, Blömer, Gage, et al. 1996). In theory, this could eliminate the need for cytokine stimulation of quiescent target cells such as HPCs during ex vivo transduction, although in practice protocols for lentiviral transduction of HPCs typically use cytokines (Cartier et al. 2009). Furthermore, the integration profile and reduced likelihood of causing clonal dominance render lentiviral vectors an attractive alternative to gammaretroviral vectors that have been implicated in insertional mutagenesis events. In contrast to SIN gammaretroviral vectors and gammaretroviral vectors with intact LTRs, SIN lentiviral vectors do not appear to affect the engraftment dynamics or repopulation properties of HPCs, and produce higher donor chimerism in peripheral tissue of mice receiving transduced long-term repopulating HPCs (Gonzalez-Murillo et al. 2008; Kustikova et al. 2009).

Rat fibroblasts, human primary macrophages and brain cells of adult rats were the first non-dividing cells stably transduced by HIV-1-based lentiviral vectors (Naldini, Blömer, Gallay, et al. 1996; Naldini, Blömer, Gage, et al. 1996). Gene
delivery to non-dividing cells in the absence of cytokine stimulation has been reported in HPCs arrested in G<sub>0</sub>, rodent liver cells, the sub-retinal space of rat eyes, pancreatic islets, airway epithelia, muscle and primary blood lymphocytes stimulated with cytokines (Naldini & Verma 2000).

**1.6.3 Modification of the HIV-1 genome to produce replication incompetent lentiviral vectors for gene therapy**

Since the biology of HIV-1 has been well-characterised, lentiviral vectors based on the HIV-1 genome have been particularly well-developed. The HIV-1 genome comprises 9 open reading frames encoding 15 proteins and cis acting sequences (Table 1.7). The development of HIV-1-based lentiviral vectors for use in gene therapy is based on modifying the HIV-1 genome (Table 1.7). The sequences that are retained include those required for efficient cellular entry, reverse transcription, nuclear uptake, chromosomal integration and transcription, while the sequences that are eliminated include those required for assembly and budding of infectious viral particles (Ailles & Naldini 2002). The need to ensure lentiviral vectors cannot produce infectious particles is also based on the high risk of danger posed by HIV-1 to humans. The risk of wild-type HIV-1 formation by recombination is minimised through the use of a 4 plasmid packaging system, with strategic deletions, minimal overlapping sequences and heterologous sequences.

Third generation lentiviral vectors with the SIN configuration have a deletion of up to 400 base pairs (bp) in the U3 region of the 3'-LTR, which is copied to the 5'-LTR during reverse transcription (Zufferey et al. 1998). These vectors are produced from 4 plasmids. The lentiviral transfer vector plasmid contains a transgene expression cassette flanked by essential cis-acting sequences, with the Rev response element (RRE)
Table 1.7: Summary of functional characteristics of HIV-1 wild type sequences and use in third generation lentiviral vectors

(Ailles & Naldini 2002)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Function of Product(s)</th>
<th>Required for cellular uptake, integration, expression, or packaging in wild-type HIV-1?</th>
<th>Deleted in Third Generation Lentiviral Vector?</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'LTR promoter region</td>
<td>Promote transcription of early HIV-1 sequences with cellular RNA Polymerase II</td>
<td>Expression</td>
<td>Yes</td>
</tr>
<tr>
<td>tat</td>
<td>Forms a complex with cellular cyclin T and Cdk9 to bind TAR and increase efficiency of transcription</td>
<td>Expression</td>
<td>Yes</td>
</tr>
<tr>
<td>rev</td>
<td>Binds RRE of unspliced viral RNA to facilitate nuclear export</td>
<td>Expression</td>
<td>Provided on a separate packaging plasmid, since only required in trans</td>
</tr>
<tr>
<td>gag</td>
<td>Produces structural proteins Gag and Gag-Pol that are packaged into virions. Gag and Gag-Pol are cleaved to produce mature matrix (MA), capsid (CA), nucleocapsid (NC) and p6 proteins by viral protease in newly budded virions</td>
<td>Packaging</td>
<td>No</td>
</tr>
<tr>
<td>env</td>
<td>Produces mature envelope glycoproteins surface (SU) and transmembrane (TM) by glycosylation and cleavage. SU binds extracellular CD4 in cellular uptake. TM assists fusion between viral and target cell membranes</td>
<td>Cellular uptake and packaging</td>
<td>Yes, except for RRE which is required in trans and so provided on a separate packaging plasmid. Another packaging plasmid contains a heterologous envelope protein such as vesicular stomatitis virus glycoprotein G, vsvg</td>
</tr>
<tr>
<td>SD and SA</td>
<td>Assist processing of viral RNA</td>
<td>Expression</td>
<td>No</td>
</tr>
<tr>
<td>Gene</td>
<td>Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ψ</td>
<td>Initiates dimerisation of RNA genome, assists interactions with Gag and packaging into virions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vif</td>
<td>Assists replication in non-permissive cells by overcoming an endogenous inhibitor, viral stability, and efficient reverse transcription</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vpr</td>
<td>Induces cell cycle arrest at G2 when 5'LTR promoter-enhancer has greatest activity, assists nuclear import</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nef</td>
<td>Assists efficient reverse transcription, promotes endocytosis of CD4 during cellular uptake</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vpu</td>
<td>Targets intracellular CD4 for proteolysis to prevent Env-CD4 complexes, facilitates viron release, downregulates cell surface expression of endogenous MHC-I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pol</td>
<td>Viral protease cleaves Pol to produce protease (PR), reverse transcriptase (RT) and integrase (IN) in newly budded virions. RT produces a DNA copy of the RNA genome for integration and then digests viral RNA. Integrase catalyses genomic integration of viral DNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cPPT (in pol) and 3PPT</td>
<td>cPPT primes synthesis of downstream DNA, 3'PPT primes synthesis of upstream DNA. Generates a 118 bp central DNA flap assists nuclear import</td>
<td></td>
<td></td>
</tr>
<tr>
<td>polyA</td>
<td>No product, but causes termination of transcription</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expression and packaging</th>
<th>Expression</th>
<th>Cellular uptake and expression</th>
<th>Cellular uptake</th>
<th>Integration</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Replaced with non-HIV-1 sequence</td>
</tr>
</tbody>
</table>
between SD and SA sites. The remaining essential trans-acting protein coding sequences are divided over 3 packaging plasmids. Sequences not essential for gene transfer are deleted, including most of env (except for RRE) and the accessory genes vif, vpr, vpu and nef (Table 1.7). Use of an internal promoter enables deletion of the transcriptional activator tat (Naldini & Verma 2000). The central polypurine tract (cPPT) is retained to promote nuclear translocation (Follenzi et al. 2000).

The 3 packaging plasmids are pMDLg/pRRE, pRSV-Rev and pMD.G (Dull et al. 1998). pMDLg/pRRE contains gag and pol driven by the human cytomegalovirus (CMV) promoter-enhancer and the RRE downstream of pol. pRSV-Rev contains rev controlled by the Rous sarcoma virus (RSV) promoter-enhancer. pMD.G encodes the heterologous envelope protein, under the control of the CMV promoter-enhancer. The G glycoprotein of the vesicular stomatitis virus (VSVG) confers broad tropism and high stability on lentiviral particles, permitting concentration by ultracentrifugation and long-term storage while further reducing the chance of recombination (Ailles & Naldini 2002).

Deletion of HIV-1 sequences reduces the likelihood of recombination. The only major components of wild type HIV-1 that remain are gag and pol, as these are essential for gene transfer (Dull et al. 1998). Their expression depends on rev expression in trans, so gag and pol cannot be expressed outside the packaging cells even in the event of recombination (Dull et al. 1998). Assaying for p24, which is cleaved from GAG, in serially passaged transduced cells can test for replication-competent lentivirus (RCL; Naldini & Verma 2000). Recombination to form the wild-type HIV-1 genome is not possible, because less than 5% of the wild type HIV-1
genome is present in the vector, and less than 25% is in the packaging plasmids (Thomas et al. 2003).

1.6.4 Existing data evaluating the safety of lentiviral vectors compared to gammaretroviral vectors

Several studies have set out to quantify the theoretical safety advantages offered by lentiviral vectors using the preclinical models described in Section 1.4.3, which purport to provide a read-out of genotoxicity. In the Luc activation assay, both the MLV-based gammaretroviral and HIV-1-based lentiviral backbones with intact LTR U3 regions led to transcriptional read-through, but the gammaretroviral backbone and not the lentiviral backbone was associated with significant enhancer activity (Hendrie et al. 2008). In the cellular immortalisation assay involving serial replating of murine bone marrow, the SIN gammaretroviral construct containing the SFFV promoter-enhancer transformed primary bone marrow, but to a lesser degree than gammaretroviral vectors with intact LTRs (Modlich et al. 2006).

The in vitro LMO2 activation assay revealed that a SIN lentiviral vector with internal cellular β-globin and EFS promoter-enhancer elements and flanking insulator elements did not activate LMO2 expression, in contrast to a gammaretroviral vector with 2 intact LTRs (Ryu et al. 2008). The hypersensitive site 4 insulator from the chicken β-globin locus (cHS4) reduced LMO2 expression several-fold in that assay and has also been shown to suppress clonal dominance in Jurkat cells, a human T cell line, when contained in both LTRs of a lentiviral vector with GFP expression controlled by the MSCV promoter (Evans-Galea et al. 2007). Finally, in the CDKN2A−/− tumour-prone mouse assay for insertional mutagenesis, it was found that gammaretroviral vectors are more genotoxic than lentiviral vectors, such that 10-fold higher integration
load of lentiviral vectors with active LTRs would be required to approach the same risk of a matched-design gammaretroviral vector (Montini et al. 2009).

A SIN lentiviral vector has been produced in large-scale clinical levels from a clone generated using concatemeric array transfection, rather than the standard 4 plasmid transfection system (Throm et al. 2009). This vector construct contains the EFS promoter, a codon-optimised IL2Rγ cDNA and the cHS4 insulator within the LTRs, which contain full deletions of promoter-enhancer sequences. In light of these developments, the US Recombinant DNA Advisory Committee approved the use of SIN lentiviral vectors for use in a US-based SCID-X1 gene therapy trial based at St Jude Children’s Research Hospital (Williams 2009), which is currently recruiting participants (ClinicalTrials.gov Identifier NCT01512888).

1.6.5 Promoter-enhancer selection to minimise trans-activation of proto-oncogenes

Selection of the internal promoter-enhancer element to use in a SIN construct is a further aspect of vector design that can be used to improve safety. A relatively inactive regulatory element with weak enhancer activity may be less likely to cause insertional activation. Studies have therefore examined the trans-activating potential of promoter-enhancer elements that differ in transcriptional activity levels.

In a plasmid-based assay testing enhancer interaction with a minimal thymidine kinase (TK) promoter controlling Luc expression, the cellular PGK and EFS promoter-enhancer elements demonstrated significantly weaker effect on TK activity than the retroviral myeloproliferative sarcoma virus and SFFV promoter-enhancer elements (Zychlinski et al. 2008). Compared to vectors with intact LTR elements, SIN
gammaretroviral vectors did not enhance Luc activity, and the PGK and EFS elements within a SIN gammaretroviral vector did not enhance TK activity above background in this system. The same study also tested these vectors in the clonal dominance insertional mutagenesis assay. The SIN gammaretroviral vector with an internal EFS promoter-enhancer did not produce any mutants with serial replating capacity compared to a SIN gammaretroviral vector with an internal MLV promoter-enhancer. Lin⁻ bone marrow cells transduced with the SIN gammaretroviral vector with internal EFS promoter did not activate EVII expression above background within 14 days after transduction, in contrast to cells transduced with the SIN gammaretroviral vector with an internal MLV promoter-enhancer, and with or without an insulator.

Another study investigated the transcriptional strength and trans-activation potential of commonly used promoter-enhancer elements, including the cytomegalovirus enhancer/chicken β-actin (CAG), EF1α, PGK, CMV immediate early, Simian virus 40 early, MoMLV LTR, myeloproliferative sarcoma virus LTR, MSCV LTR, Friend SFFV, MoMLV SIN LTR, HIV SIN LTR and FV SIN LTR promoter-enhancer elements (Weber & Cannon 2007). The CAG and EF1α promoter-enhancers led to the highest levels of transgene expression across different cell types, including lymphoid cells. A cell specific pattern of trans-activation was observed, with the EF1α and PGK promoter-enhancers displaying the highest trans-activation of a 1.0-kb downstream minimal thymidine kinase promoter-element in K562 cells, respectively, and no trans-activation above background in Jurkat cells.

A few studies have examined levels of therapeutic efficacy achieved with different cellular promoters in murine models of SCID-X1 gene therapy. The enhancer-less ubiquitously acting chromatin opening element (UCOE) lacks classic enhancer
activity but has been shown to drive greater transgene expression in bone marrow and peripheral blood than the strong SFFV and CMV promoter-enhancers, reconstitute immunity in a mouse model of SCID-X1 gene therapy, and confer greater resistance to methylation-mediated silencing (Zhang et al. 2007; Zhang et al. 2010). SIN HIV-1-based lentiviral vectors containing the PGK promoter-enhancer and a 1.1-kb fragment of the endogenous \( IL2R_\gamma \) proximal promoter-enhancer restored immunity in \( \gamma_c \)-deficient mice and did not demonstrate the bias towards integration near proto-oncogenes observed with a SFFV-containing SIN lentiviral vector (Huston et al. 2011).

SIN HIV-1-based lentiviral vectors containing the EFS promoter-enhancer driving a codon-optimised \( IL2R_\gamma \) and a 1.2-kb \( IL2R_\gamma \) proximal promoter-enhancer driving the genomic \( IL2R_\gamma \) sequence have also been reported (Zhou et al. 2010). Both vectors achieved phenotype correction in a murine model of SCID-X1 and enabled T cell development \textit{in vitro} following transduction of CD34\(^+\) cells from SCID-X1 patients. The EFS-containing vector did not activate \( LMO2 \) expression in the Jurkat cell-based \( LMO2 \)-activation assay (Section 1.4.3), and was not associated with tumourigenesis in mice. Consequently, the EFS-containing lentiviral vector has been selected for testing in a phase I clinical trial being conducted by the team at St Jude Children’s Research Hospital.

\textbf{1.6.6 Development of genome editing and vectors targeting ‘safe harbours’}

While lentiviral vectors represent a promising new generation of retrovirus-based integrating vectors for gene therapy with possible safety advantages over gammaretroviral vectors, a number of alternative viral and non-viral vectors are being developed that may offer further reduced genotoxicity risks. While both
gammaretroviral and lentiviral vectors have been consistently shown to target genes, vectors based on FVs display a more random pattern of genomic integration and do not appear to favour coding regions (Trobridge et al. 2006). SIN FV-based vectors displayed the least trans-activation and transcriptional read-through in the in vitro Luc activation assay (Hendrie et al. 2008). Sustained correction of canine leukocyte adhesion deficiency, a haematopoietic disorder, has been achieved with FV vectors, with an absence of genotoxic complications and vector targeting of oncoviruses compared to gammaretroviral vectors (Bauer et al. 2008). Canine pyruvate kinase deficiency has also been corrected using an FV-based vector and an in vivo expansion strategy (Trobridge et al. 2012).

Since insertional mutagenesis is associated with the use of integrating viral vectors, non-viral vectors that permit stable expression of IL2Rγ may be beneficial. The key challenge faced by these systems, however, is achieving the levels of therapeutic efficacy demonstrated with integrating viral vectors. Given the low efficiency of non-viral strategies, there is a need for these developments to be combined with advances in cellular biology, in order to expand rare genetically modified cells. Such advances would include an improved ability to expand CD34+ HPCs and maintain the CD34+ phenotype during culturing. Non-viral vectors with mitotically stable transgene expression, such as episomal vectors and artificial chromosomes, are being developed but at present these approaches are limited by poor efficiency. Since applications of gene therapy that target HPCs involve many rounds of cell division, a key challenge for episomal vectors is ensuring that they can replicate and segregate following cell division.
The φC31 integrase system has a more predictable integration profile than the semi-random integration of retroviral vectors. The φC31 integrase is encoded by a Streptomyces bacteriophage and pairs a phage attP site with a non-identical attB site in the Streptomyces genome, and has been shown to mediate integration in mammalian cells (Groth et al. 2000; Calos 2006). The φC31 integrase can direct integration of a plasmid encoding IL2Rγ and neo' and containing the attB sequences towards intergenic regions on chromosomes 13 and 18 in Jurkat cells and IL2Rγ-null ED40515(-) cells (Ishikawa et al. 2006). φC31 integrase-based gene delivery has not achieved phenotype correction in a mouse model of disease, and it is unknown whether the need to expand and select transfected cells for G418 resistance could produce somatic genetic lesions which may lead to reduced safety.

In a similar system, the integration profile of Sleeping Beauty transposonase is paired with the efficient cell and nuclear entry of integrase-deficient HIV-1 lentiviral vectors (Vink et al. 2009). This has been used to target integration of an expression cassette away from transcriptionally active genes. The development of site-specific integration using fusion proteins comprising HIV-1 integrase and zinc-finger proteins that recognise specific DNA motifs may be possible (Su et al. 2009).

Endogenous mechanisms of homologous recombination can be exploited to achieve gene correction. Gene correction of defective IL2Rγ has made progress in overcoming the poor inefficiency associated with such strategies (Urnov et al. 2005). A mutation hot-spot in the fifth exon of IL2Rγ was targeted by the DNA-binding domain of 2 engineered zinc finger proteins, each with 4 zinc finger motifs. Since 1 α-helix of a zinc finger motif can recognise 3 bp of DNA, the 2 engineered proteins could recognise a 24 bp sequence within IL2Rγ that occurred uniquely in the genome. Attached to this
DNA-binding domain was a DNA nuclease, which induced a double-stranded DNA break. The correct form of the gene was supplied to the cell on a plasmid, and homologous recombination mechanisms inserted the correction into the genome. Eighteen per cent of co-transfected cells contained corrected \textit{IL2Rγ} and inheritance was mitotically stable. More recently, \textit{IL2Rγ} has been corrected in CD34\(^+\) cord blood progenitor cells and Epstein-Barr virus-transformed B cells (Lombardo et al. 2007). Alternative gene correction strategies based on synthetic transcription activator-like effector nucleases (TALENs) are also being developed (Miller et al. 2011).

Gene correction strategies avoid the risk of insertional mutagenesis associated with integrating vectors, however, the risks associated with this process have not yet been investigated in a comprehensive manner. For example, the frequency of off-target events, such as DNA breaks occurring at loci other than \textit{IL2Rγ}, needs to be assessed. Ultimately, animal studies would be needed to assess the efficacy and safety of gene correction technologies and clinical trials are likely to be distant. Given the promising safety and efficacy outcomes achieved with redesigned SIN gammaretroviral and lentiviral vectors in preclinical models of SCID-X1 gene therapy, viral-based gene therapy is more likely to provide more immediate therapy for SCID-X1 patients lacking an HLA-identical sibling donor.

1.7 RISK ASSESSMENT AND PERCEPTIONS OF ACCEPTABLE RISK LEVELS IN GENE THERAPY TRIALS

The development of a new medical treatment, such as a redesigned vector for SCID-X1 gene therapy, requires a combination of biological and ethical analyses, and these analyses are interrelated (Figure 1.3). The question of whether the development of a new treatment is justified depends on the severity of disease and the availability and
Figure 1.3: The interdependence between biological and ethical analyses of safety and efficacy in translational gene therapy. Preclinical development of new treatments, such as gene therapy, depends on ethical analysis of the disease context (considerations listed to the left of the blue arrow above the words “ethical analyses”). Biological analyses lead to the optimal design for a gene therapy vector and preclinical assessment of safety and efficacy in the best available preclinical model (considerations listed to the left of the blue arrow below the words “biological analyses”). Given the limitations of preclinical models, however, the likely safety and efficacy outcomes in humans remain uncertain (considerations to the right of the purple arrow below the words “biological analyses”). Ethical analysis of the disease context also influences the degree of uncertainty that is acceptable when a trial is initiated (consideration to the right of the purple arrow above the words “ethical analyses”).
limitations of existing treatments. Interpreting biological assessments of the safety and efficacy of a new treatment depends on an understanding of the limitations of preclinical model systems and the uncertainties which may remain after preclinical tests. The question of whether to proceed to a clinical trial of the new treatment depends on how much certainty about these risks is required for making such decisions, and how these uncertain risks weigh against the known risks associated with the existing treatments.

1.7.1 Uncertainty of risk in translational research

The development of leukaemia in 5 of 20 patients treated in the otherwise highly successful French and British trials of gene therapy for SCID-X1 highlights the degree of scientific and clinical uncertainty in translational research and the difficulties of risk assessment. As discussed in Section 1.4, the frequency of insertional mutagenesis manifesting in leukaemia was not anticipated on the basis of existing preclinical data. Significant progress has been made with respect to understanding the risk of insertional mutagenesis and characterising vector integration behaviour (Section 1.5), and the design of new vectors, including the SIN gammaretroviral vector being tested in the international SCID-X1 gene therapy trial and the SIN lentiviral vector being tested at St Jude Children’s Research Hospital (Sections 1.6.1–1.6.5).

Notwithstanding this progress, inevitably a degree of uncertainty with regards to the risk of insertional mutagenesis will remain when these vectors are tested in humans for the first time. Preclinical models of SCID-X1 gene therapy do not recapitulate the human phenotype perfectly (Section 1.4.2), and there is ultimately no substitute for human clinical trials. Since insertional mutagenesis events can take years to present, it
will be many years until definitive conclusions can be drawn about the safety of these redesigned vectors. Phase I clinical trials inevitably, and by definition, involve both a degree of risk and persistent uncertainty, since safety and efficacy cannot be known before the trial is initiated. Decisions regarding the commencement of such studies, particularly where the risks may be severe or life-threatening, are therefore informed by both scientific judgment of the likelihood of risk and moral judgments about the seriousness, meaning and implications of these risks.

Part of this challenge concerns the assessment of the ‘acceptable risks’ with respect to the potential benefits, and part concerns who is responsible for making these decisions about when phase I trials are initiated (Deakin et al. 2009). Trials of SCID-X1 gene therapy are illustrative of the challenges that confront translational research in not only gene therapy but also in many other areas of emergent research. A critical discourse about risk in clinical research may assist with addressing the challenge of assessing uncertain risks in translational gene therapy research.

1.7.2 Defining acceptable levels of risk and uncertainty in phase I clinical trials of gene therapy

One of the most profound, and intractable, challenges of risk assessment is to strike the most appropriate balance between ‘accepting’ and minimising risks. Key questions about the acceptability of risk in clinical trials include who makes decisions about risk, whether the possible benefits justify the level of risk, what values underlie these decisions, how these decisions are scrutinised, and what kinds of questions should be asked to both protect research subjects and allow research to progress (Deakin et al. 2009). Such judgments are informed, but not determined by, preclinical in vitro and in
vivo testing and involve epistemological questions about the extent of legitimate extrapolation of preclinical data to humans.

By acknowledging that some uncertainty is irreducible, researchers must consider the level of certainty about the potential risks and benefits that both they and the research participants require when deciding about initiating and continuing a study, and about participating in it, respectively (Deakin et al. 2010). Researchers who understand the limitations of models in preclinical studies may accept that the safety and efficacy data generated in those systems do not necessarily resolve safety concerns, but may nevertheless suggest that clinical trials are justified by the burden of illness and the absence of other therapeutic alternatives. Similarly, participants, and parents of paediatric participants, may not only appreciate the potential benefits of such research, but may also understand that absolute certainty is not possible in the absence of clinical trials, and may accept qualitative estimates of risk expressed in terms such as ‘unlikely’, ‘uncommon’ and ‘most’. Evaluating the degree of certainty that researchers and participants require in order to make decisions is therefore both an empirical question concerning the predictive capacities of preclinical models, and a moral question regarding the degree of certainty required to justify trials in humans.

The benefits associated with a potential life-long cure may be an additional consideration that may influence acceptable levels of risk in gene therapy trials. For example, gene therapy approaches that purport to provide a life-long cure for infants with diseases that would otherwise require costly treatments at regular intervals, such as enzyme replacement therapy for ADA-SCID, may justify a higher level of risk in clinical trials, as well as being associated with longer-term economic benefits.
1.7.3 Potential influence of perceptions about risk during clinical trial decision-making

Risk can be defined as the probability that a particular adverse event will occur within a defined period of time, where that probability is a function of hazard and exposure (Chicken & Posner 1998). However, despite the appeal of definitions suggesting that risk can be objectively defined and quantified, it is also clear that risk encompasses subjective elements (Garland 2003). Consequently, different understandings of risk have evolved in many disciplines, including statistics, psychology, economics, philosophy, law and the social sciences. A number of socio-demographic factors are recognised to influence risk perception, including age, sex, culture and education (Slovic 2000; Renn & Rohrmann 2000; Dosman et al. 2001; Glendon et al. 1996). It is also well-established that risk is perceived differently by lay people and experts in that field (Slovic et al. 1980; Slovic 1987). Consequently, what may appear to a researcher as an unacceptably high level of risk may appear as an only hope to an anxious patient with dire prospects. Such patients may not see a choice if the alternative is death. Or, conversely, what may appear to a researcher as an insignificant risk may cause a patient more concern. Similarly, patients and researchers may weigh the prospect of benefit differently.

Knowledge about the type, extent and frequency of risks is crucial for managing risks. Preclinical safety analyses can assist with measuring risks, understood in their objective and probabilistic sense. However, when potential risks associated with a new therapy are balanced with the potential benefits, assessments of risks involve both scientific and normative judgment, in that they draw on values that are personal,
cultural and moral (Deakin et al. 2009). Identifying what these values might be and articulating how they can affect decision-making may facilitate future decision-making.

1.7.4 Potential influence of the clinical context of risk assessment

The level of risk that may be considered ‘acceptable’ during clinical trial decision-making may be determined by the perceived ‘need’ for, or attraction of, a new treatment, and this in part reflects the decision-maker’s knowledge and experience, including the severity of the disease in question, life expectancy, expected quality of life, and the availability and limitations of existing treatment alternatives. An increased level of risk may therefore be justifiable, and acceptable, in clinical trials of gene therapy for life-threatening conditions for which treatments do not exist or are inadequate, such as aggressive paediatric brain tumours and the severe neonatal phenotype of ornithine transcarbamylase (OTC) deficiency. Conversely, it may be more difficult to justify increased levels of risk for conditions that are not life-threatening, such as Leber’s congenital amaurosis (LCA). Although diseases such as SCID-X1, ADA-SCID, CGD and adrenoleukodystrophy (ALD) can be treated by BMT, the associated effectiveness and risks are centre-dependent and are limited by the availability of compatible donors (Section 1.2.2).

At present, little is known about how the clinical context and perceptions of the need for a new treatment impact on assessments of risks when decisions are made about clinical trials.
1.7.5 Communication of uncertain risks to potential research participants

Risk assessment by researchers in the process of designing a clinical trial represents only one dimension of the risk assessments that occur during clinical research. Ultimately it is the participants in research who bear the risks of research and these individuals, or their parents in trials involving children, need to make informed assessments of risks and potential benefits before providing consent to research. It is crucial, therefore, that researchers are aware of the manner in which participants understand risks, the types of information they require and the degree of certainty they are satisfied with in order to make decisions about trials. This is both because this informs trial design and because it facilitates effective risk communication. This, in turn, requires that researchers are aware of approaches to risk communication, and are willing to consider the strategies that might best empower individual participants and address their specific informational needs (Visschers et al. 2009; Epstein et al. 2004; Paling 2003).

In relation to gene therapy research in particular, the focus should be on communicating the uncertainty of risk information, such as the meaning of preclinical risk estimates, the potential limitations of such data and the knowledge gaps that exist. It is also essential that communication strategies clearly and unambiguously convey that the primary purpose of research is to generate knowledge, rather than to provide direct therapeutic benefit (Appelbaum et al. 1987; Appelbaum et al. 1982). And because gene therapy research frequently involves children, researchers need to consider not only the best approach to use to communicate the details of the study to parents and other surrogate decision-makers, but also strategies to empower children to
achieve an appropriate level of participation in decision-making (Wendler 2008; Kunin 1997).

In general, methods of communicating risk that are more personalised and interactive lead to greater levels of knowledge and understanding, and facilitate more informed decision-making with respect to risks (Trevena et al. 2006). Question prompt lists (QPLs) are an example of a communication tool that may increase participants’ knowledge and encourage them to take an active role in dialogue about risk (Brehaut et al. 2010; Hazen et al. 2010). For example, a QPL written for patients, or parents of paediatric patients, who are considering enrolment in a clinical trial of gene therapy targeting the bone marrow compartment, such as the new SCID-X1 gene therapy trials, could direct them to inquire about: (i) risk and survival rates following transplantation at local centres; (ii) the prognosis of participants in different hematopoietic gene therapy trials, particularly the number and status of patients that developed leukaemia in the SCID-X1 gene therapy trials; and (iii) whether there are any differences between the proposed vector and the vector used in the SCID-X1 trial, including possible safety developments. Research participants who understand the rationale and design of a gene therapy trial, including the known and unknown risks and benefits, may have greater satisfaction with uncertain risk information.
General methods used throughout the laboratory component of this thesis are detailed in this chapter. Results Chapters 3, 4 and 5 contain additional methods used specifically in those chapters. A list of all solutions and buffers that are referred to throughout Chapters 2 to 4 are provided in Tables 2.1, 2.3 and 2.4. All general methods described in this chapter were performed personally by the author.

2.1 REAGENTS

All chemicals and reagents used were of analytical grade. Solutions were prepared in MilliQ water and sterilised by autoclaving at 121°C for 20 minutes when required, unless stated otherwise. Autoclaved MilliQ water was used in molecular biology reactions unless stated otherwise.
2.2 CULTURING OF MAMMALIAN CELLS

2.2.1 Plasticware, buffers and solutions

All cells were cultured at 37°C with humidified air containing 5% CO₂. Plasticware, including tissue culture flasks and serological pipettes, were obtained from BD Biosciences (San Jose, USA). Tissue culture media was filter sterilised using a 0.22 μm cellulose acetate membrane (Corning Incorporated, Acton, USA).

2.2.2 Passaging of mammalian cells

Adherent HEK-293 cells (ATCC Number CRL-1573; Graham et al. 1977) were cultured in Complete DMEM (Table 2.1). Cells growing in a 75 cm² flask were dislodged using 0.25% Trypsin-EDTA (Table 2.1), and were split 1:10 every 3–4 days by mechanically dislodging the cells from the flask surface in 10 mL fresh Complete DMEM, and adding 1 mL of cells to 10 mL of fresh Complete DMEM in a new flask.

Non-adherent ED-7R cells (Kumaki et al. 1999) were cultured in Complete RPMI (Table 2.1). Cells growing in a 75 cm² flask were split 1:3 by centrifuging the cells at 230 × g at room temperature (RT; 25°C) for 5 minutes, resuspending the cell pellet in 3 mL fresh Complete RPMI, and adding 1 mL of cells to approximately 10 mL fresh Complete RPMI in a new flask.
### Table 2.1: Buffers and solutions used for tissue culture

<table>
<thead>
<tr>
<th>Solution</th>
<th>Composition</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone marrow cryomedia</td>
<td>RPMI 1640 Medium (Life Technologies, Carlsbad, USA) supplemented with 40% (v/v) foetal bovine serum (SAFC Biosciences, Lenexa, USA) and 10% (v/v) dimethyl sulphoxide (DMSO; Sigma-Aldrich, St Louis, USA) (filter-sterilised)</td>
<td>Cryopreservation of murine bone marrow (Section 3.3.2.2)</td>
</tr>
<tr>
<td>Buffer (i)</td>
<td>Phosphate buffered saline (MP Biomedicals LLC, Solon, USA) supplemented with 1% (v/v) heat-inactivated foetal bovine serum (SAFC Biosciences) and 0.4% (w/v) sodium azide (Ajax Finechem, Sydney, Australia)</td>
<td>Antibody-labelling of cells for flow cytometry (Section 3.3.2.3)</td>
</tr>
<tr>
<td>2 M Calcium chloride</td>
<td>2 M calcium chloride (Amresco, Solon, USA)</td>
<td>Calcium phosphate-mediated transfection of HEK-293 cells (Section 3.3.1.2)</td>
</tr>
<tr>
<td>Complete DMEM</td>
<td>Dulbecco’s Modified Eagle’s Medium supplemented with 10% (v/v) foetal bovine serum (SAFC Biosciences,) (filter-sterilised)</td>
<td>Culturing of HEK-293 cells (Section 2.2.2)</td>
</tr>
<tr>
<td>Complete RPMI</td>
<td>RPMI 1640 Medium (Life Technologies) supplemented with 10% (v/v) foetal bovine serum (SAFC Biosciences) (filter-sterilised)</td>
<td>Culturing of non-ED-7R cells (Section 2.2.2)</td>
</tr>
<tr>
<td>DMEM cryomedia</td>
<td>Dulbecco’s Modified Eagle’s Medium supplemented with 40% (v/v) foetal bovine serum (SAFC Biosciences) and 10% (v/v) DMSO (Sigma-Aldrich, St Louis, USA) (filter-sterilised)</td>
<td>Cryopreservation of HEK-293 cells (Section 2.2.5)</td>
</tr>
<tr>
<td>Foetal bovine serum (FBS; SAFC Biosciences, Lenexa, USA)</td>
<td>Heat-inactivated (50ºC for 60 minutes)</td>
<td>Routine culturing of mammalian cells (Section 2.2)</td>
</tr>
</tbody>
</table>
### Chapter 2: Materials and methods

<table>
<thead>
<tr>
<th><strong>2× Hepes-buffered saline (2× HBS)</strong></th>
<th>280 mM sodium chloride (Amresco) and 50 mM N-2-Hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES; Biochemicals, Gymea, Australia); pH 7.10</th>
<th>Calcium phosphate-mediated transfection of HEK-293 cells (Section 3.3.1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phosphate-buffered saline without calcium and magnesium (PBS -/-; MP Biomedicals LLC, Solon, USA)</strong></td>
<td>1.06 mM potassium phosphate monobasic, 155.17 mM sodium chloride, 2.97 mM sodium phosphate dibasic; pH 7.4</td>
<td>Preparation of a single cell suspension (Section 2.2.3); antibody-labelling of cells for flow cytometry (Section 3.3.2.3)</td>
</tr>
<tr>
<td><strong>Red cell lysis buffer</strong></td>
<td>150 mM ammonium chloride (BDH Laboratory Supplies, Poole, UK), 10 mM potassium hydrogen carbonate (BDH Laboratory Supplies), 1.25 mM ethylenediaminetetraacetic acid (EDTA; Amresco); pH 7.3</td>
<td>Lysis of red blood cells in bone marrow and blood samples (Sections 3.3.2.2 and 3.3.2.3)</td>
</tr>
<tr>
<td><strong>RPMI cryomedia</strong></td>
<td>RPMI 1640 Medium (Life Technologies) supplemented with 40% (v/v) FBS (SAFC Biosciences) and 10% (v/v) DMSO (Sigma-Aldrich) (filter-sterilised)</td>
<td>Cryopreservation of ED-7R cells (Section 2.2.5)</td>
</tr>
<tr>
<td><strong>Sodium monohydrogen phosphate</strong></td>
<td>0.15 M sodium monohydrogen phosphate (Merck, Darmstadt, Germany)</td>
<td>Calcium phosphate-mediated transfection of HEK-293 cells (Section 3.3.1.2)</td>
</tr>
<tr>
<td><strong>Splenocyte cryomedia</strong></td>
<td>90% (v/v) FBS (SAFC Biosciences) and 10% (v/v) DMSO (Sigma-Aldrich) (filter-sterilised)</td>
<td>Cryopreservation of murine splenocytes (Section 3.3.3)</td>
</tr>
<tr>
<td><strong>Sucrose cushion</strong></td>
<td>20% (w/v) sucrose (BDH-Merck, Kilsyth, Australia) in phosphate-buffered saline (without calcium and magnesium)</td>
<td>Concentration of lentiviral vector particles (Section 3.3.1.3)</td>
</tr>
<tr>
<td><strong>10% TE</strong></td>
<td>10 mM tris-chloride (Amresco and Merck), 1 mM ethylenediaminetetraacetic acid (EDTA; Amresco); pH 8</td>
<td>Calcium phosphate-mediated transfection of HEK-293 cells (Section 3.3.1.2)</td>
</tr>
</tbody>
</table>
2.2.2 Preparation of a single cell suspension

Adherent cells were trypsinised to create a single cell suspension. Cells were washed in 5 mL phosphate buffered saline (without calcium and magnesium; PBS-/-), and incubated with 2 mL trypsin/ethylenediaminetetraacetic acid (EDTA; GIBCO; Table 2.1) at 37°C for 5 minutes. Trypsin was inactivated by the addition of approximately 10 mL complete DMEM. A single cell suspension of non-adherent cells was prepared mechanically by pipetting cells up and down.

2.2.3 Estimating cell density

Cell density was estimated by counting the number of viable cells in 0.1 μL covering a 0.1 mm³ surface on an improved Neubauer haemocytometer (Neubauer, Weber, England). Non-viable cells were excluded on the basis of Trypan Blue uptake (0.4%; Sigma-Aldrich, St Louis, USA), which was added to an equal volume of cells before counting.

2.2.4 Long-term storage

Two to 3 million HEK-293 or ED-7R cells were frozen for long-term storage in 1 mL of DMEM cryomedia or RPMI cryomedia (Table 2.1), respectively, in 1.8 mL
cryovials (Nalge Nunc International, Rochester, USA). Cells were frozen overnight at -80°C inside foam insulation before long-term storage in liquid nitrogen. Cells were thawed from long-term storage in a water bath at 37°C for 1–2 minutes, and added to 9 mL of media to dilute the dimethyl sulphoxide (DMSO; Sigma-Aldrich). Cells were pelleted (230 × g at RT for 5 minutes), resuspended in fresh media, and placed in a 25 cm² tissue culture flask.

2.3 MOLECULAR BIOLOGY TECHNIQUES

2.3.1 Bacterial strains and growth media

Chemically competent HB101 (Promega, Madison, USA) and JM109 (Promega) and XL10-Gold Ultracompetent (Agilent Technologies, Santa Clara, USA) strains of *Escherichia coli* were used in the chemical transformation of bacteria (Table 2.2). The electrocompetent Stop Unwanted Recombination Events (SURE; Agilent Technologies) strain of *E. coli* were used in transformations by electroporation (Table 2.2).

All *E. coli* strains were grown in Luria-Bertani broth (LB; Table 2.3), with the exception of SURE cells, which were grown in Super Optimal Broth with Carbolite repression (SOC; Table 2.3). LB and SOC contained ampicillin (100 μg/mL; Roche Diagnostics Australia Pty Ltd, Castle Hill, NSW) for selection of bacteria with plasmids encoding β-lactamase. LB plates contained ampicillin and 1.5% (w/v) bacto-agar (BD Biosciences, Sparks, USA).
Table 2.2: Genotypes of bacterial strains

<table>
<thead>
<tr>
<th>Bacterial strain</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>HB101 (Promega)</td>
<td>F–, thi-1, hsdS20 (ρB, mB), supE44, recA13, ara-14, leuB6, proA2, lacY1, galK2, rpsL20 (str′), xyl-5, mtl-1</td>
</tr>
<tr>
<td>JM109 (Promega)</td>
<td>endA1, recA1, gyrA96, thi, hsdR17 (ρ−, m−), relA1, supE44, Δ(lac-proAB), [F′ traD36, proAB, laqI6ΔAM15]</td>
</tr>
<tr>
<td>XL10-Gold (Agilent Technologies)</td>
<td>Tet′, Δ(mcrA)183, Δ(mcrCB-hsdSMR-mrr)173, endA1, supE44, thi-1, recA1, gyrA96, relA1, lac, Hte, [F′ proAB, laqI6 ΔAM15, Tn10, (Tet′)]</td>
</tr>
<tr>
<td>SURE (Agilent Technologies)</td>
<td>e14′ (McrA′), Δ(mcrCB-hsdSMR-mrr)171, endA1, gyrA96, thi-1, supE44, relA1, lac, recB, recJ, sbcC, umuC::Tn5 (Kan′), uvrC, [F′ proAB, laqI6 ΔAM15, Tn10, (Tet′)]</td>
</tr>
</tbody>
</table>

Table 2.3: Liquid and solid media used for bacterial work

<table>
<thead>
<tr>
<th>Solution</th>
<th>Composition</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB agar</td>
<td>Luria-Bertani broth supplemented with 1.5% (w/v) bacto-agar (BD Biosciences, Sparks, USA)</td>
<td>Culturing of <em>E. coli</em> bacterial strains (Section 2.3.1)</td>
</tr>
<tr>
<td>10×Luria-Bertani broth (LB)</td>
<td>10% (w/v) sodium chloride (Amresco), 10% (w/v) bacto-yeast extract (BD Biosciences), 5% (w/v) bacto-tryptone (BD Biosciences); pH 7</td>
<td>1× LB used for culturing of <em>E. coli</em> bacterial strains (Section 2.3.1)</td>
</tr>
<tr>
<td>Super Optimal Broth with Carbolite repression (SOC)</td>
<td>2% (w/v) bacto-tryptone (BD Biosciences), 0.5% (w/v) bacto-yeast extract (BD Biosciences), 10 mM sodium chloride (Amresco, Solon, USA), 2.5 mM potassium chloride (Ajax Finechem, Taren Point, Australia), 10mM magnesium chloride (Ajax Finechem) and 20 mM glucose (BDH Laboratory Supplies, Poole, UK) (all components autoclave-sterilised except glucose, which was filter-sterilised)</td>
<td>Culturing of <em>E. coli</em> bacterial strains (Section 2.3.1)</td>
</tr>
</tbody>
</table>
2.3.2 Nucleic acid isolation

2.3.2.1 Small-scale isolation of plasmid DNA

Small-scale quantities of plasmid DNA (up to 20 μg) were isolated from bacterial cells by alkaline lysis, using a QIAprep Miniprep Kit (Qiagen GmbH, Hilden, Germany), according to the manufacturer’s protocol. An overnight culture was prepared by inoculating 3 mL LB with cells from a well-isolated colony growing on an LB agar plate, and incubating these cells overnight at 37°C with shaking. Pelleted bacterial cells from the overnight culture were resuspended in resuspension buffer, lysed by the addition of lysis buffer containing sodium hydroxide and then neutralised. This lysate was loaded onto a silica-gel membrane, which selectively bound plasmid DNA in high salt buffer, permitting removal of RNA, proteins and metabolites during washing stages. Plasmid DNA was eluted by the addition of low salt buffer and stored at -20°C.

2.3.2.2 Large-scale isolation of plasmid DNA

A NucleoBond Plasmid Maxi Kit (Machery-Nagel, Düren, Germany) was used to isolate larger quantities of plasmid DNA (up to 500 μg) for transfections. A 3 mL starter culture was used to inoculate 500 mL LB, which was incubated overnight at 37°C with shaking. Glycerol stocks of the overnight culture were prepared by mixing 0.5 mL glycerol with 1 mL of bacterial cell suspension, which were stored at -80°C. Pelleted bacterial cells from the 500 mL overnight culture were lysed under alkaline conditions. This lysate was neutralised, cleared by centrifugation and loaded in low salt buffer onto an anion-exchange resin in a column, which operated by gravity flow. The resin was washed to remove RNA and proteins. Plasmid DNA was eluted in high salt buffer, and further purified by precipitation with isopropanol (Merck, Darmstadt,
Germany) and washing in 70% (v/v) ethanol (Merck) to remove salt, before resuspension of the pellet in 10% tris-ethylenediaminetetraacetic acid (10% TE; Table 2.4) and long-term storage at -20°C.

2.3.2.3 Isolation of genomic DNA from cultured cells

Genomic DNA was isolated from mammalian cells using a QIAamp DNA Blood Mini Kit (Qiagen), following the protocol supplied by the manufacturer. Pelleted cells were lysed in the presence of Qiagen Protease and RNase A, incubated at 56°C for 10 minutes, and loaded onto a QIAamp Spin Column. Genomic DNA adsorbed to the column membrane in the presence of ethanol, enabling proteins and RNA to be removed by washing before purified genomic DNA was eluted in low salt buffer. Once isolated, genomic DNA was stored at 4°C to avoid freeze-thaw induced shearing.

2.3.2.4 Isolation of genomic DNA from mouse tissue

Genomic DNA was isolated from mouse tissue using the Gentra Puregene Blood Kit (Qiagen), following the manufacturer’s instructions. Pelleted cells were lysed using a solution that contained an anionic detergent and a DNA stabilising agent. Samples were RNase-treated and proteins were removed by salt precipitation. Genomic DNA was precipitated with isopropanol, washed, dissolved in 10% TE and stored at 4°C.
### Table 2.4: Buffers and solutions used for molecular biology

<table>
<thead>
<tr>
<th>Solution</th>
<th>Composition</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>10× annealing buffer</td>
<td>1.375 M tris-chloride (Amresco, Solon, USA and Merck, Darmstadt, Germany), 500 mM magnesium chloride (Ajax Finechem, Taren Point, Australia); pH 7.4</td>
<td>Annealing of complementary single-stranded oligonucleotides (Sections 3.3.4.2 and 4.3.11)</td>
</tr>
<tr>
<td>6× loading dye</td>
<td>50% (w/v) sucrose (BDH-Merck, Kilsyth, Australia), 0.05% (w/v) bromophenol blue (Amresco), 0.5% (w/v) xylene cyanol (Amresco)</td>
<td>Agarose gel electrophoresis (Section 2.332)</td>
</tr>
<tr>
<td>3 M sodium acetate</td>
<td>3 M sodium acetate (BDH Laboratory Supplies, Poole, UK); pH 5.2</td>
<td>Ethanol precipitation (Section 2.3.3.6)</td>
</tr>
<tr>
<td>TE</td>
<td>100 mM tris-chloride (Amresco and Merck), 10 mM ethylenediaminetetraacetic acid (EDTA; Amresco); pH ranges used 7.0-8.5</td>
<td>Solvent for nucleic acids (Section 2.3)</td>
</tr>
<tr>
<td>10% TE</td>
<td>10 mM tris-chloride (Amresco and Merck), 1 mM EDTA (Amresco); pH 8</td>
<td>Solvent for nucleic acids (Section 2.3);</td>
</tr>
<tr>
<td>50×Tris-acetate/EDTA</td>
<td>2 M tris-acetate (Amresco and Merck), 50 mM EDTA (Amresco), pH 8.0</td>
<td>1× TAE used in agarose gel electrophoresis of DNA (Section 2.332)</td>
</tr>
<tr>
<td>(TAE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5×Tris-borate/EDTA</td>
<td>450 mM tris-borate (Amresco and Merck), 10 mM EDTA (Amresco), pH 8</td>
<td>1× TBE used in acrylamide gel electrophoresis of DNA (Section 2.333)</td>
</tr>
<tr>
<td>(TBE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 mM Tris-chloride</td>
<td>10mM tris-chloride (Amresco and Merck), pH 8.5</td>
<td>Solvent for nucleic acids (Section 2.3)</td>
</tr>
</tbody>
</table>
2.3.2.5 Isolation of RNA

Total RNA was isolated from mammalian cells, including mouse tissue, using an RNeasy Mini Kit (Qiagen), according to the manufacturer’s protocol. Pelleted cells which had been stored at -80°C were resuspended in lysis buffer containing guanidine isothiocyanate to inactivate RNases, and were homogenised in a QIAshredder Column, which shears genomic DNA. The sample was loaded onto an RNeasy Spin Column to allow total RNA to bind to the column membrane, which was treated with RNase-free DNase to cleave genomic DNA. Impurities were removed from the membrane during washing stages before RNA was eluted into RNase-free water, and stored at -80ºC.

Total RNA was isolated from mouse blood using Trizol (Life Technologies, Carlsbad, USA). Five hundred microlitres of Trizol reagent, which contained phenol and guanidine isothiocyanate, were added to each 200 μL blood sample to lyse cells. The addition of chloroform (Chem-Supply, Gillman, Australia) and subsequent centrifugation separated samples into an aqueous phase, containing RNA, and an organic phase, which was discarded. RNA was precipitated with isopropanol, washed, dissolved in RNase-free water, and stored at -80ºC.

2.3.3 Routine nucleic acid procedures

2.3.3.1 Estimation of nucleic acid concentration and purity

Nucleic acid concentration in solution was estimated by measuring absorbance at 260 nm in using a Nanodrop 1000 (Thermo Fisher Scientific, Wilmington, USA), assuming a solution of double stranded DNA at 1 mg/mL has an absorbance of 20 at 260 nm. The purity of DNA and RNA was estimated by calculating the ratio of absorbance at 260
nm to absorbance at 280 nm, which is 1.8 for pure nucleic acid samples (Sambrook, Fritsch, & Maniatis, 1989).

2.3.3.2 Agarose gel electrophoresis of DNA

DNA fragments were electrophoresed in horizontal agarose gels, in 1× tris-acetate/EDTA electrophoresis buffer (TAE; Table 2.4). Agarose (Amresco) was dissolved in TAE, and ethidium bromide (1 μg/mL; Amresco) was added. The percentage agarose content in these gels depended upon the expected size of the fragment to be visualised, and ranged from 0.8% (w/v) for larger DNA fragments to 2.5% (w/v) for smaller DNA fragments. Loading buffer (Table 2.4) was added to samples prior to loading. The DNA size standards used included Hyperladders I, IV and V (Bioline, London, UK), GeneRuler Low Range DNA Ladder (Thermo Fisher Scientific), Marker III (λ DNA digested with EcoRI and HindIII; Roche Diagnostics) and DNA Marker VIII (pUCBM21 digested with HpaII and pUCBM21 digested with DraI and HindIII; Roche Diagnostics). Samples were electrophoresed at a field strength of approximately 11 V/cm. DNA fragments in agarose gels were visualised and photographed using a NovaLine (NovaLine, Encinitas, USA) or AlphaImager Gel Documentation System (ProteinSimple, Santa Clara, USA).

2.3.3.3 Polyacrylamide gel electrophoresis of DNA

Polyacrylamide gel electrophoresis was used when finer resolution of relatively small DNA fragments was required. DNA fragments were electrophoresed in vertical gels containing 10%–15% acrylamide (Bio-Rad Laboratories, Hercules, USA) and 1× trisborate/EDTA electrophoresis buffer (TBE; Table 2.4). Polymerisation was induced by the addition of a 1% volume of ammonium persulfate (10% (w/v) in water; Sigma-
Chapter 2: Materials and methods

Aldrich) and a 0.05% volume of N,N,N',N'-tetramethylethylenediamine (TEMED; Sigma-Aldrich). Ten per cent polyacrylamide gels were used for DNA fragments of approximately 15 bp to 400 bp, while 15% polyacrylamide gels were used for DNA fragments of less than 200 bp. DNA fragments separated by polyacrylamide gel electrophoresis were stained using SYBR Gold (Life Technologies) and visualised using a Typhoon Trio (GE Healthcare, Buckinghamshire, UK).

2.3.3.4 Column-based DNA purification

DNA was purified from gel slices or following enzymatic reactions, such as PCR, using a Wizard SV Gel and PCR Clean-Up Kit (Promega), as per the manufacturer’s protocol. Gel slices containing DNA fragments of interest were melted in a high salt buffer, which disrupts hydrogen bonds within the agarose polymer, and which also contained guanidine isothiocyanate. Alternatively, the high salt buffer was added to samples containing enzymatically-modified DNA. The DNA and high salt buffer solution was loaded onto a resin containing silica-gel particles which selectively bound the DNA fragment in the presence of high salt. Unbound impurities, such as agarose, proteins, ethidium bromide and salts, were removed by washing, and the purified DNA fragment was eluted in 10 mM tris-chloride (Table 2.4) for storage at -20ºC.

2.3.3.5 Phenol:chloroform extraction of DNA

Phenol-chloroform extractions were performed to purify DNA from proteins following enzymatic reactions, such as restriction endonuclease digestion. The phenol-chloroform extraction method relies upon the separation following centrifugation of an organic phase, containing proteins, and an aqueous phase, containing nucleic acids (Sambrook et al. 1989). An equal volume of phenol:chloroform:isoamylalcohol (25:24:1; saturated
with 10 mM tris, 1 mM EDTA, pH 8.0; Sigma-Aldrich) was added to samples before briefly vortexing and centrifugation at 15,000 × g for 15 minutes. The organic phase, containing denatured proteins, was removed. The organic extraction was repeated by adding an equal volume of chloroform:isoamylalcohol (24:1), inverting and centrifuging at 15,000 × g for 15 minutes. DNA was purified by ethanol precipitation (Section 2.3.3.6).

2.3.3.6 Ethanol precipitation of DNA

DNA was purified from salts or following phenol-chloroform extraction by precipitation with ethanol (Sambrook et al. 1989). Ten per cent 3M sodium acetate (pH 5.2; Table 2.4; BDH Laboratory Supplies, Poole, UK) and 2 new volumes of ethanol, respectively, were added, as well as 2-3 μL of glycogen (20 mg/mL in water; Roche Diagnostics) to assist with visualising the DNA pellet. Samples were incubated at -80°C for 1-3 hours or at -20°C overnight. The precipitated DNA was pelleted by centrifugation at 15,000 × g for 60 minutes at 4°C. The DNA pellet was washed using 80% (v/v) ethanol and resuspended in the desired volume of water, TE or 10 mM tris-chloride (Table 2.4).

2.3.4 Transformation of competent bacteria with plasmid DNA

2.3.4.1 Bacterial transformation using the heat-shock method

Competent HB101 and JM109 *E. coli* strains were transformed with bacterial plasmids according to the manufacturer’s protocol (Promega). Briefly, 100 μL of cells were mixed with 1-50 ng plasmid DNA, incubated on ice for 10 minutes, heat-shocked at
42°C for 30 seconds, incubated on ice for a further 2 minutes, and then incubated in 1 mL LB at 37°C with shaking for 1 hour.

Ultracompetent XL10-Gold cells were transformed with bacterial plasmids as described by the manufacturer (Stratagene). These bacterial cells have reduced size bias and are transformed by large plasmids and smaller plasmids equally, and so were used in transformations with large lentiviral transfer plasmid constructs. Briefly, 50 μL of bacterial cells were incubated on ice for 10 minutes, mixed with approximately 400 ng plasmid DNA in the presence of β-mercaptoethanol, incubated on ice for 30 minutes, heat-shocked at 42°C for 30 seconds, incubated on ice for a further 2 minutes, and finally incubated at 37°C with shaking in 1 mL LB for 1 hour.

One hundred microlitres of transformed bacterial cells and the remaining transformed cells, pelleted and resuspended in 100 μL LB, were spread on LB plates and incubated overnight at 37°C.

2.3.4.2 Bacterial transformation using electroporation

Electrocompetent SURE cells were transformed with plasmid DNA according to the manufacturer’s protocol. Briefly, 30 μL of cells were added to approximately 50-100 pg of plasmid DNA on ice, and then transferred to a pre-chilled electroporation cuvette (0.1 cm gap; Bio-Rad Laboratories). Cells were pulsed using 1700V in a Bio-Rad electroporator with resistance set at 200 ohms and capacitance set at 25 μF, and then 960 μL of pre-warmed SOC was added. Cells were incubated at 37°C with shaking for 1 hour.
Transformed bacterial cells were diluted 1:100, and 5 μL, 10 μL, 20 μL, 50 μL and 100 μL of the diluted cells were spread on LB plates and incubated overnight at 37°C.

2.3.5 Recombinant DNA techniques

2.3.5.1 Restriction endonuclease digestion of DNA

DNA was digested for 1-8 hours with 5-10 U per μg DNA of restriction endonuclease (typically purchased from New England Biolabs or Roche Diagnostics) at the optimised temperature in the presence of the appropriate buffer, as recommended by the manufacturer of the enzyme. For double digests, 1 suitable buffer was selected for the reaction.

2.3.5.2 Ligation of DNA molecules

Ligation reactions were performed using T4 DNA Ligase (Promega). For ligations of vector and insert DNA molecules, reactions were typically performed in a total volume of 10 μL containing 1.5 U of ligase, reaction buffer, 1 mM ATP, 50 ng of vector DNA and a 10-fold molar excess of insert DNA. Ligations of a linker cassette to genomic DNA were performed in a minimal volume (10-40 μL) under the same conditions, but with over 1000-fold molar excess of linker cassette (Sections 3.3.4.2 and 3.3.5.2). Reactions were incubated at room temperature for 1 hour or at 16°C or 4°C overnight. T4 DNA Ligase was inactivated the following day by incubating the reaction mixture at 70°C for 10 minutes.
2.3.6 Subcloning of PCR products

PCR products, such as the final A-tailed PCR products of integration site methodologies (Sections 3.3.4.2 and 3.3.5.4), were subcloned for sequencing using the TOPO-TA Cloning Kit (Life Technologies). The protocol provided by the manufacturer was followed, with the exception that the ligation reaction volume was halved. A typical reaction was conducted in a final volume of 3 μL, containing 0.5 μL of the TOPO-TA vector and 2 μL of PCR product. The whole reaction was used to transform either the One Shot TOP10 cells provided in the kit or HB101 cells (Section 2.3.4.1).

2.3.7 Sanger sequencing of DNA

Fragments of DNA were sequenced using an AB 3730xl instrument (Australian Genome Research Facility, Sydney, Australia). When plasmid DNA samples were sequenced, a 12 μL reaction contained 600-1500 ng of DNA and 9.6 pmol of sequencing primer. When PCR products were directly sequenced without subcloning, a 12 μL reaction contained 3-12 ng of DNA and 9.6 pmol of sequencing primer. The sequence chromatograms in the sequencing trace files were visualised using Sequence Scanner Version 2 (Life Technologies).
3.1 INTRODUCTION

HIV-1-based lentiviral vectors are promising tools for use in applications of gene therapy targeting the haematopoietic compartment, such as gene therapy for SCID-X1. In addition to the safety advantage of deleted LTR promoter-enhancer elements, the SIN configuration of third generation lentiviral vectors allows an opportunity for careful selection of an internal cellular promoter-enhancer element to control γc expression. Given concerns about the risk of insertional mutagenesis, and the potential for enhancer activity within regulatory elements to trans-activate genes near the vector integration, the selection of an appropriate promoter-enhancer element is a critical issue.
in gene therapy for SCID-X1. On the one hand, the promoter-enhancer element must be sufficiently active to enable sufficient levels of \( \gamma c \) expression to restore \( \gamma c \)-dependent IL signalling. On the other hand, promoter-enhancer elements with strong enhancer activity, such as the viral CMV and SFFV promoter-enhancer elements are to be avoided, as these may be more likely to trans-activate neighbouring genes.

In addition, an ideal promoter-enhancer element would enable physiological levels of \( \gamma c \) expression in the correct target cell population, and preferably at the stages of lymphocyte ontogeny in which \( \gamma c \) expression is required. For this reason, promoter-enhancer elements isolated from housekeeping genes are commonly used. Examples of cellular promoter-enhancer elements that have been used in preclinical SCID-X1 gene therapy studies include a 1,205 bp fragment of the human endogenous IL2R\( \gamma \) proximal promoter-enhancer, a 1,177 bp human EF1\( \alpha \) promoter-enhancer fragment (EF1\( \alpha \) promoter-enhancer), a 242 bp human EFS promoter-enhancer fragment, a 516 bp human PGK promoter-enhancer fragment (PGK promoter-enhancer), a 481 bp human Wiskott-Aldrich syndrome protein promoter-enhancer fragment (WASP promoter-enhancer), and a 2.2 kb enhancer-less human UCOE (Zychlinski et al. 2008; Ginn et al. 2010; Huston et al. 2011; Zhang et al. 2007; Thornhill et al. 2008; Zhou et al. 2010). The EF1\( \alpha \) and EFS promoter enhancers were both isolated as regulatory elements of EF1\( \alpha \).

An additional matter to consider during the selection of an internal promoter-enhancer element is the level of \( \gamma c \) expression necessary to restore \( \gamma c \)-dependent IL signalling required for both T cell and NK cell ontogeny. Previous research in the Gene Therapy Research Unit (GTRU) has contributed to understanding this area, by investigating the mutation of a patient with an atypical TNK\(^+\) SCID-X1 phenotype.
The mutation was identified as a novel splice site mutation that produced low levels of correctly-spliced γc transcript. Further investigation showed that when γc expression is limiting, signalling via the IL-15 receptor, required for NK cell ontogeny, is preferentially retained over signalling via the IL-7 receptor, required for T cell ontogeny (Smyth et al. 2007). This implies that as the transcriptional activity of the promoter-enhancer element used to control γc expression in a gene transfer vector is decreased, reconstitution of the T cell compartment is more likely to be impaired. The selection of an ideal internal promoter-enhancer element to include within a SIN lentiviral vector for SCID-X1 gene therapy thus involves a balancing act. The internal promoter-enhancer element must be sufficiently transcriptionally active to enable robust reconstitution of both the T and NK cell compartments. At the same time, the enhancer activity of the promoter-enhancer element should be minimal in order to minimise the risk of insertional mutagenesis via trans-activation.

A panel of third generation SIN lentiviral vectors with different internal promoter-enhancer elements controlling γc expression have been developed by the GTRU and tested for preclinical safety and efficacy in a mouse model of SCID-X1 (Smyth et al. 2007; Ginn et al. 2010). The elements used include the EF1α, PGK and WASP promoter-enhancers. The EF1α promoter-enhancer is approximately 5.0-fold and 8.8-fold more transcriptionally active than the PGK and WASP promoter-enhancers, respectively, in human lymphoid IL2Rγ-null ED-7R cells (Smyth et al. 2007). The EF1α promoter-enhancer element is approximately 3–4-fold more transcriptionally active than the PGK promoter-enhancer element in human CD34+ cells (Woods et al. 2000; Salmon et al. 2000; Ramezani et al. 2000). The enhancer activity of these elements is unknown.
Chapter 3: Evaluating the safety and efficacy of lentiviral vectors in vivo

T cell and NK cell reconstitution in mice treated with an EF1α.γc vector were equivalent to wild-type levels, and reconstituted T cells could respond to mitogen stimulation (Ginn et al. 2010). In mice treated with the PGK.γc vector, however, T cell reconstitution was incomplete. Interestingly, the proportion of reconstituted NK cells in mice treated with the PGK.γc vector was similar to wild-type. There was no reconstitution of T cells or NK cells in mice treated with the WASP.γc vector. In addition to these analyses of immune reconstitution, a cohort of 14 mice treated with the EF1α.γc vector was maintained long-term to monitor for safety. Four of the mice developed lymphomas 5–9 months post-treatment, in contrast to an absence of lymphomas in 25 mice transplanted with wild-type Sca1⁺ bone marrow (BM). Extensive molecular analyses eliminated insertional mutagenesis or γc over-expression as possible causes.

Although the mice treated with wild-type Sca1⁺ BM received an equivalent cell dose to the EF1α.γc-treated mice, 100% of the wild-type cells were expressing γc at physiological levels. By contrast, it was estimated that approximately 22 cent of cells received by the EF1α.γc-treated mice were transduced, and within that subset it was unknown what proportion of cells were expressing sufficient γc to restore γc-dependent signalling, since γc expression can be influenced by the genomic location of the proviral integration (Cassani et al. 2009). Thus, the 2 cohorts received significantly different doses of cells that were competent to reconstitute the lymphocyte compartment. Given that the overall levels of lymphocyte reconstitution were equivalent between the 2 cohorts, it was hypothesised that the Sca1⁺ BM transduced with EF1α.γc were under greater pressure to proliferate in order to reconstitute the lymphocyte compartment. This replicative stress may have led to an acquisition of somatic genetic lesions and ultimately malignancy. The tumours observed in the
Chapter 3: Evaluating the safety and efficacy of lentiviral vectors in vivo

EF1α.γc-treated mice were hypothesised to have been caused by replicative stress, since insertional mutagenesis had been excluded as a causative mechanism.

Replicative stress is, at present, poorly understood as a potential risk factor in haematopoietic gene therapy, and ongoing research in the GTRU is investigating the impact of transplantation of limiting doses of wild-type Sca1⁺ BM in IL2Rγ-deficient mice (GTRU, unpublished). Analyses of patient samples from the French SCID-X1 trial revealed that the clonal complexity of reconstituted T cells correlates positively with the dose of transplanted γc⁺CD34⁺ cells, however, it remains unknown whether lower cell doses are associated with replicative stress (Wang et al. 2010). At present, it is unknown to what extent features of vector design may reduce the effective dose of HPCs competent for lymphocyte reconstitution and whether this can increase the likelihood of malignancy. For example, conventional thinking in the gene therapy field would favour the selection of a less transcriptionally active promoter-enhancer element with low enhancer activity to control γc expression, in order to minimise the risk of insertional mutagenesis. However, the proportion of transduced HPCs expressing sufficient γc levels to reconstitute the lymphocyte compartment may be reduced if such a promoter-enhancer is used, and this may be associated with replicative stress. Thus, despite the potential safety advantages of using a less transcriptionally active promoter-enhancer element to reduce the risk of insertional mutagenesis, there is a potential safety disadvantage if this may lead to replicative stress.

Given the potential implications of replicative stress for vector design, the aim of this chapter was to investigate the effect of limiting γc expression through the selection of a promoter-enhancer element with reduced transcriptional activity on the frequency of lymphomagenesis caused by replicative stress. This chapter also
investigates whether the use of a less transcriptionally active promoter-enhancer element is associated with reduced complexity of reconstituted lymphocytes, using lentiviral vector integration sites as a unique mark for individual clones. Lower levels of lymphocyte reconstitution were observed in mice treated using the PGK-γc vector compared to the EF1α-γc vector, consistent with previous observations for those vectors (Ginn et al. 2010). However, no difference was observed between the cohorts of mice treated with either of these vectors in the frequency of lymphomagenesis. Analyses of the degree of clonal complexity of reconstituted lymphocytes were inconclusive. Therefore, the question of whether a less transcriptionally active promoter-enhancer is associated with reduced clonal complexity was not definitively answered.

3.2 CHAPTER 3 HYPOTHESIS AND AIMS

The hypotheses underlying the investigations in this chapter were:

(1) An internal promoter-enhancer element with reduced transcriptional activity within a SIN IL2Rγ-encoding lentiviral vector leads to reduced levels of lymphocyte reconstitution;

(2) Reduced transcriptional activity of an internal promoter-enhancer element correlates with reduced clonal complexity of reconstituted lymphocytes; and

(3) Reduced transcriptional activity of an internal promoter-enhancer element within a SIN IL2Rγ-encoding lentiviral vector correlates with a higher incidence of lymphomagenesis.

The aims of this chapter were:
(1) To set up cohorts of mice treated with the EF1α.γc and PGK.γc lentiviral vectors;
(2) To compare the levels of lymphocyte reconstitution achieved with the EF1α.γc and PGK.γc vectors;
(3) To monitor cohorts of mice treated with the EF1α.γc and PGK.γc vectors long-term (15 months) for the development of haematological malignancies;
(4) To investigate whether insertional mutagenesis contributed to the development of any haematological malignancies observed; and
(5) To measure the degree of clonal complexity in reconstituted lymphocytes of mice treated with the EF1α.γc and PGK.γc vectors by analysis of unique integration sites.

3.3 MATERIALS AND METHODS

All work described in this chapter was performed personally by the author, except where otherwise is indicated.

3.3.1 Lentiviral vector production

3.3.1.1 Plasmid constructs used in lentiviral vector production

The production of third-generation HIV-1-based VSVG-pseudotyped lentiviral vectors using a transfer vector plasmid and the packaging plasmids pMDLg/pRRE, pRSV-Rev and pMD.G has been described (Dull et al. 1998). The lentiviral transfer vector plasmids used during this project, pEF1α.γc and pPGK.γc (Table 3.1), have been reported (Smyth et al. 2007; Ginn et al. 2010), and were based on pRRLsin.cPPT.hCMV.EGFP.WPRE (Follenzi et al. 2000). Packaging plasmids,
pMDLg/pRRE, pRSV-Rev and pMD.G were used in addition to the lentiviral transfer vector plasmids during lentiviral vector production (Table 3.1).

**Table 3.1: Plasmids used for lentiviral vector production**

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Size (bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pEF1α.γc</td>
<td>Lentiviral transfer vector plasmid based on pRRLsin.cPPT.CMV.EGFP.WPRE</td>
<td>9261</td>
<td>Ginn <em>et al.</em> 2010</td>
</tr>
<tr>
<td></td>
<td>(Follenzi <em>et al.</em> 2000). Expression cassette consists of a 1,177 bp human</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EF1α promoter-enhancer element and the IL2Rγ transgene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pPGK.γc</td>
<td>Lentiviral transfer vector plasmid based on pRRLsin.cPPT.CMV.EGFP.WPRE</td>
<td>8582</td>
<td>Ginn <em>et al.</em> 2010</td>
</tr>
<tr>
<td></td>
<td>(Follenzi <em>et al.</em> 2000). Expression cassette consists of a 516 bp human</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PGK promoter-enhancer element and the IL2Rγ transgene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pMDLg/pRRE</td>
<td>Packaging plasmid containing gag and pol controlled by a cytomegalovirus</td>
<td>8895</td>
<td>Dull <em>et al.</em> 1998</td>
</tr>
<tr>
<td></td>
<td>(CMV) promoter-enhancer element and the Rev response element (RRE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>downstream of pol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pRSV-Rev</td>
<td>Regulatory plasmid with rev expression controlled by a Rous sarcoma virus</td>
<td>4174</td>
<td>Dull <em>et al.</em> 1998</td>
</tr>
<tr>
<td></td>
<td>(RSV) promoter-enhancer element</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pMD.G</td>
<td>Envelope plasmid containing vesg, the heterologous envelope G glycoprotein</td>
<td>5824</td>
<td>Naldini, Blömer, Gallay, <em>et al.</em></td>
</tr>
<tr>
<td></td>
<td>from the vesicular stomatitis virus, under the control of a CMV promoter-</td>
<td></td>
<td>1996; Naldini, Blömer, Gallay,</td>
</tr>
<tr>
<td></td>
<td>enhancer element</td>
<td></td>
<td><em>et al.</em> 1996</td>
</tr>
</tbody>
</table>

**3.3.1.2 Transient co-transfection of lentiviral packaging and transfer vector plasmids**

Packaging plasmids pMDLg/pRRE, pRSV-Rev and pMD.G were transiently co-transfected with lentiviral transfer vector plasmids by calcium phosphate
co-precipitation (Gorman 1985). In a typical preparation of each vector, 60 10 cm² dishes were seeded with $5 \times 10^6$ HEK-293 cells the day before transfection such that cells would be 80–90% confluent on the day of transfection.

On the day of transfection, the media was changed 3–4 hours before transfection. The following solutions were prepared for 4 dishes in a total volume of 2 mL. Solution A contained sodium monohydrogen phosphate (0.15 μM) made up to 2 mL in 2× Hepes-buffered saline (2× HBS; Table 2.1). Solution B contained a total of 160 μg of DNA and calcium chloride (0.26 mM) made up to 2 mL using 10% TE (Table 2.1); the calcium chloride was added last. The 160 μg of DNA in Solution B contained the lentiviral transfer vector plasmid, either pEF1α.γc or pPGK.γc, and the 3 packaging plasmids in an equimolar ratio.

Solution B was added to Solution A drop-wise, swirling every 2–3 drops, to give a total volume of 4 mL for 4 dishes. The precipitation reaction was incubated at RT for 20 minutes and then added to the cells. After approximately 1 hour the precipitate could be observed scattered over the cells. Cells were incubated overnight, and the media was replaced the following morning.

3.3.1.3 Lentiviral vector harvest and concentration

Supernatant containing lentiviral vector particles was harvested from HEK-293 cells co-transfected with the packaging and transfer vector plasmids at 48 and 72 hours after transfection, filtered through a 0.45 μm cellulose acetate syringe filter to remove cells (Thermo Fisher Scientific), and stored at -80°C, or at 4°C if concentrated the following day. Vector supernatant was concentrated using 2 rounds of ultracentrifugation. For the first round of ultracentrifugation, aliquots of approximately 37 mL vector supernatant
were added to 1.5 mL of 20% (w/v) sucrose (Table 2.1) in Ultraclear tubes for the SW32 rotor of the Optima L-100XP Ultracentrifuge (Beckman Coulter, Pasadena, USA). Lentiviral vector particles were pelleted at 72,000 × g at 4°C for 2 hours. Supernatant was discarded and pellets were resuspended in 1 mL PBS-/- for a second round of ultracentrifugation.

The total volume of the resuspended pellets, typically 24 mL, was divided over 2 Ultraclear tubes for the SW41 rotor of the ultracentrifuge. The 20% (w/v) sucrose solution was used to make up the remaining volume needed to fill the tubes. Lentiviral vector particles were pelleted at 72,000 × g at 4°C for 2 hours. Supernatant was discarded and each pellet was resuspended in 50–100 μL PBS-/-.

The resuspended pellets were pooled, pipetted up and down to facilitate resuspension, and allowed to further resuspend by incubation on ice overnight. Aliquots of concentrated lentiviral vector were stored at -80°C and thawed no more than twice to minimise titre loss.

### 3.3.1.4 Assignment of transducing titre to lentiviral stock

Lentiviral transducing titre was defined as the number of transducing units (TU) per mL, which estimates the number of integrated proviruses and hence functional vector particles. The number of vector integrations was determined by qPCR (Sastry et al. 2002), a method that is independent of transgene expression. Different volumes of vector supernatant (50 μL, 100 μL and 200 μL for unconcentrated supernatant; 0.5 μL, 1 μL and 2 μL for concentrated supernatant) were added to HEK-293 cells seeded at 5 \times 10^5 cells per well in a 6-well plate in 1 mL of media in the presence of Polybrene (8 μg/mL; Sigma-Aldrich). Genomic DNA was extracted (Section 2.3.2.3) for use as template DNA for qPCR.
Proviral copy numbers were quantified using Platinum Taq (5 U/μL; Life Technologies), DNA standards of known copy number, primers FPLV2 and RPLV2 which annealed to the lentiviral packaging signal, and the FAM-BHQ dual-labelled DNA probe LV2 (Table 3.2). A total volume of 25 μL contained 200 ng template DNA, reaction buffer, magnesium chloride (2mM), the 4 dNTPs (1 mM), 8 nmol of primers FPLV2 and RPLV2 and 5 nmol of probe LV2. Reactions were performed in a RotorGene 6000 (Corbett Life Science, Mortlake, Australia) and cycling consisted of a denaturing step (95°C for 10 minutes), followed by 40 cycles of template denaturing (95°C for 15 seconds) and primer annealing and extension (60°C for 60 seconds).

Table 3.2: Oligonucleotide sequences used for assigning titre to lentiviral vector stocks

<table>
<thead>
<tr>
<th>Oligonucleotidea</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPLV2</td>
<td>5’-ACCTGAAAGCGAAAGGGAAAC-3’</td>
</tr>
<tr>
<td>RPLV2</td>
<td>5’-CACCCATCTCTCTCTCTCTAGCC-3’</td>
</tr>
<tr>
<td>LV2</td>
<td>5’[FAM]-AGCTCTCGACACAGGACTCGGC-3’[BHQ]</td>
</tr>
<tr>
<td>LC-Top</td>
<td>5’-GACCCGGGAGATCTGAATTCAGTGGCACACGAGCTTA GG-3’</td>
</tr>
<tr>
<td>LC-Tsp-bottom</td>
<td>5’- AATTCCTAACTGCTGTGCCACTGAATTCAGATC-3’</td>
</tr>
<tr>
<td>CDL TRI</td>
<td>5’- CTCCCAACGAAGACAAGATCTGC-3’</td>
</tr>
<tr>
<td>LTR II</td>
<td>5’-AGCTTGCTTGTGCTTCA-3’</td>
</tr>
<tr>
<td>LTR III</td>
<td>5’-AGTATGGTCAGCTGTGCTTTG-3’</td>
</tr>
<tr>
<td>LCI</td>
<td>5’- GACCCGGGAGATCTGAATTC-3’</td>
</tr>
<tr>
<td>LCII</td>
<td>5’- AGTGGCAGACAGTTAGG-3’</td>
</tr>
</tbody>
</table>

aAll oligonucleotides were purchased from Sigma-Aldrich.
3.3.2 Treatment of γc-deficient mice with genetically-modified Sca1⁺ bone marrow cells

3.3.2.1 Mouse strains used for the murine model of SCID-X1 gene therapy

*IL2Rγ⁻/⁻* mice and *IL2Rγ⁻/⁻Rag2⁻/⁻* mice were used in this study. *IL2Rγ⁻/⁻* mice (Cao et al. 1995) on a C57Bl/6 background were purchased from The Jackson Laboratory (Bar Harbor, ME, USA) and provided the donor bone marrow cells that were genetically modified using *IL2Rγ*-encoding lentiviral vectors. *IL2Rγ⁻/⁻Rag2⁻/⁻* mice, also on a C57Bl/6 background, were kindly provided by Dr Stephen Nutt (The Walter and Eliza Hall Institute, Melbourne, Australia) and were used as recipients for genetically modified *IL2Rγ⁻/⁻* bone marrow cells. These mice were deficient in *Rag2* as well as *IL2Rγ*, which was necessary because approximately 10% of peripheral blood mononuclear cells in *IL2Rγ⁻/⁻* mice are CD3⁺ T cells due to leaky T cell ontogeny. Wild-type C57Bl/6 mice, which provided wild-type bone marrow cells for transplantation, were purchased from the Animal Resources Centre (ARC; Perth, Australia).

Mice were housed in micro-isolator cages in the Bioservices Facility of the Children’s Medical Research Institute under specific pathogen free (SPF) conditions. All research was conducted with ethics approval obtained from the Animal Care and Ethics Committee of the Children’s Medical Research Institute and The Children’s Hospital at Westmead (ACEC Project Number C218).

3.3.2.2 Isolation and transduction of *IL2Rγ⁻/⁻* Sca1⁺ bone marrow cells for transplantation into *IL2Rγ⁻/⁻Rag2⁻/⁻* recipient mice

The murine model for SCID-X1 gene therapy involving lentiviral transduction of *IL2Rγ⁻/⁻* HPCs and transplantation into *IL2Rγ⁻/⁻Rag2⁻/⁻* recipient mice has been
described previously (Figure 3.1; Ginn et al. 2010). Bone marrow was flushed from the femurs and tibias of 7 to 10 weeks old IL2Rγ−/− and wild-type donor mice using RPMI (Life Technologies) and a 26G needle. Bone marrow cell suspensions were filtered through a 70 μm cell strainer and red cells were lysed using red cell lysis buffer (Table 2.1). Cells were washed using RPMI, counted, and cryopreserved in bone marrow cryomedia (Table 2.1; Sections 2.2.4–2.2.5).

Sca1+ BM cells were enriched for using the EasySep SCA1 positive selection kit (StemCell Technologies, Vancouver, BC, Canada). These cells were seeded at 1 × 10^6 cells/mL in StemSpan SFEM serum-free media (StemCell Technologies), which had been supplemented with 50 U/mL penicillin (Life Technologies), 50 mg/mL streptomycin (Life Technologies), 100 ng/mL mSCF (R&D Systems, Minneapolis, USA), 100 ng/mL hFlt-3 (R&D Systems), 100 ng/mL hIL-11 (R&D Systems), and 20 ng/mL mIL-3 (R&D Systems). Typically, 1 × 10^6 cells in 1 mL of media were cultured in a single well of a 24-well plate. Sca1+ BM cells were transduced overnight (approximately 16 hours) with EF1α.γc or PGK.γc lentiviral vectors at an MOI of 100. At approximately 28 and 4 hours before transplantation, IL2Rγ−/−Rag2−/− recipient mice were given sub-lethal irradiation by a Gammasell 3000 Elan irradiator using an open Cs^{137} source (MDS Nordion, Ottawa, Canada), administered as a split dose (3.1 Gy at 28 hours and 1.6 Gy at 4 hours before transplantation). The cells that had been exposed to lentiviral vectors were washed and resuspended in RPMI for intravenous injection into the tail veins of the irradiated IL2Rγ−/−Rag2−/− recipient mice.
Figure 3.1: Schematic diagram of mouse model for SCID-X1 gene therapy. Bone marrow was isolated from IL2Rγ-/- donor mice and Sca1+ haematopoietic progenitor cells are enriched for by positive selection of Sca1+ cells. Sca1+ cells are transduced overnight with either of the EF1α.γc or PGK.γc lentiviral vectors (MOI = 100). Cells are washed, and a dose of 5 x 10^5 vector-exposed cells are injected into the tail veins of IL2Rγ-/-Rag2-/- recipient mice, which have been conditioned with a split sub-lethal dose of radiation (4.7 cGy). EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; PGK, phosphoglycerate kinase; MOI, multiplicity of infection.
3.3.2.3 Evaluation of immune reconstitution by flow cytometric analysis of lymphocyte markers

Two hundred microlitres of peripheral blood were collected from the tail veins of treated \( IL2R\gamma^{--}Rag2^{--} \) mice into heparinised micro-haematocrit tubes (Hirschmann Laborgeräte, Eberstadt, Germany), and then transferred into 1.5 mL Eppendorf tubes containing 20 μL of heparin sodium (5000 IU in 5 mL, diluted 1 in 10 in PBS; Pfizer, New York, USA). Anti-coagulated blood was added to 10 mL red cell lysis buffer (Table 2.1) to lyse red blood cells. Cells were then pelleted at 5,000 × g and resuspended in 50 μL of buffer (i) (Table 2.1) for labelling with 0.5 μL of the following antibodies: CD2-fluorescein isothiocyanate (FITC), CD45-allophycocyanin (APC), CD3-AlexaFluor647, CD3-phycoerythrin (PE), B220-APC-cyanine7, NK1.1-PE, CD4-FITC and CD8-APC (Table 3.3). After 30–60 minutes of antibody staining on ice, cells were washed in 2 mL PBS/-, pelleted at 5,000 × g and resuspended in 200 μL of PBS/- for fluorescence-activated cell sorting (FACS) analysis using a FACSCanto instrument (BD Biosciences) and FACSDiva software (Version 6.1.3). Lymphocytes were gated based on their forward and side scatter profile and expression of CD2 as a lymphocyte marker. A minimum of 10,000 gated events were acquired, where possible.

To estimate total white cell counts, 10 μL of peripheral blood were collected into 5 μL heparin. Red blood cells were lysed using 45 μL of 2% (v/v) acetic acid and white blood cells were counted using a haemocytometer.

3.3.3 Macroscopic examination and analysis of cellular characteristics of murine lymphomas

Treated \( IL2R\gamma^{--}Rag2^{--} \) mice that appeared to develop lymphoma were sacrificed following a cardiac puncture to obtain up to 1 mL of peripheral blood. Lymphoid tissue
Table 3.3: Antibodies used in flow cytometry analysis of immune reconstitution

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Supplier</th>
<th>Used as marker for</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD2-fluorescein isothiocyanate (FITC)</td>
<td>BD Biosciences</td>
<td>Lymphocytes</td>
</tr>
<tr>
<td>CD45-allophycocyanin (APC)</td>
<td>BD Biosciences</td>
<td>Lymphocytes and monocytes</td>
</tr>
<tr>
<td>CD3-AlexaFluor647</td>
<td>BD Biosciences</td>
<td>T cells</td>
</tr>
<tr>
<td>CD3-phycoerythrin (PE)</td>
<td>BD Biosciences</td>
<td>T cells</td>
</tr>
<tr>
<td>B220-allophycocyanin-cyanine7 (APC-Cy7)</td>
<td>BD Biosciences</td>
<td>B cells</td>
</tr>
<tr>
<td>NK1.1-phycoerythrin (PE)</td>
<td>BD Biosciences</td>
<td>NK cells</td>
</tr>
<tr>
<td>CD4-fluorescein isothiocyanate (FITC)</td>
<td>BD Biosciences</td>
<td>Helper T cells</td>
</tr>
<tr>
<td>CD8-allophycocyanin (APC)</td>
<td>Biolegend</td>
<td>Cytotoxic T cells</td>
</tr>
</tbody>
</table>

and infiltrated tissue, including spleen, thymus, liver, kidney and lungs, were collected and weighed. Single cell suspensions of spleen, thymus, liver and lungs in RPMI media were prepared and filtered through a 70 μm cell strainer. BM was flushed from the femur and tibia bones using RPMI and also filtered through a 70 μm cell strainer. Red blood cells in blood and single cell suspensions of tissue samples were lysed using 10 mL of red cell lysis buffer (Table 2.1), washed using RPMI and cells were counted for estimation of cell density and yields (Section 2.2.4). Samples were divided to allow 1–3 million cells for FACS analysis of lymphocyte markers (as per Section 3.2.2.3), 1–10 million cells for genomic DNA extraction (Section 2.3.2.4), and a pellet of 1–10 million cells for RNA extraction (Section 2.3.2.5). Fifteen to thirty million cells were cryopreserved using splenocyte or bone marrow cryomedia (Table 2.1; Section 2.2.5), where possible.
A small portion of spleen, liver, lung and kidney tissue was set aside before single cell suspensions were prepared for histological analysis. These samples were fixed in 10% (v/v) formalin (Sigma-Aldrich) for paraffin-embedding, 5 μm sectioning and haematoxylin and eosin (H & E) staining by the histology service at the Westmead Millennium Institute (Sydney, Australia). H & E-stained sections were visualised by light microscopy (Zeiss Axio Imager, Leica Microsystems, Wetzlar, Germany).

3.3.4 Analysis of molecular characteristics of murine lymphomas

For all analyses, the tissue source with the highest proportion of tumour clonality, as assessed by FACS analysis of immunophenotype, was selected for molecular characterisation.

3.3.4.1 qPCR analysis of vector copy number

The number of integrated lentiviral vector copies, or vector copy number (VCN), was assessed by qPCR (as per Section 3.2.1.4).

3.3.4.2 Integration site analysis by conventional ligation-mediated PCR (LM-PCR)

Lentiviral integration sites in tumour samples were identified by LM-PCR, as previously described (Mueller & Wold 1989; Kustikova et al. 2008). Genomic DNA (1.5 μg) was digested with Tsp509I for 5 hours according to the manufacturer’s instruction, and digestion was confirmed by agarose gel electrophoresis. Genomic DNA fragments were ethanol precipitated (Section 2.3.3.6) and resuspended in 13.5 μL water for use as template in a primer extension reaction, using the biotinylated LTR-binding primer CDLTRI (Table 3.4). The 50 μL reaction contained 5 U Taq DNA polymerase
Chapter 3: Evaluating the safety and efficacy of lentiviral vectors in vivo

DNA was purified from salts and enzymes (Section 2.3.3.4), and primer-extended fragments were isolated using Dynal streptavadin beads (Life Technologies) according to the manufacturer’s instructions. A linker cassette with Tsp509I-compatible ends was prepared by heating oligonucleotides LC-Top and LC-Tsp-bottom (Table 3.4) at 95°C for 5 minutes in a heating block in annealing buffer (Table 2.4), then allowing the sample to cool slowly to room temperature. The primer-extended fragments were ligated with an approximately 2,400-fold molar excess of linker cassette at 16°C or 4°C overnight. After heat-inactivation, the ligation product was used as template for an initial PCR reaction. A 50 μL reaction contained 2.5 U Taq DNA polymerase (Roche), reaction buffer, the 4 dNTPs (0.2 mM) and 25 pmol each of forward and reverse

Table 3.4: Oligonucleotide sequences used for conventional LM-PCR

<table>
<thead>
<tr>
<th>Oligonucleotide&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDLTRI</td>
<td>5’-[Biotin]CTCCCAACGAAGACAAGATCTGC-3’</td>
</tr>
<tr>
<td>LC-Top</td>
<td>5’-GACCCGGGAGATCTGAATTCTGACGAGCAGCAGTTA GG-3’</td>
</tr>
<tr>
<td>LC-Tsp-Bottom</td>
<td>5’-AATTCCTAACTGCTGCCTGAATTCTCAGATC-3’</td>
</tr>
<tr>
<td>LTRII</td>
<td>5’-AGCTTGCCCTTGAGTGCTCCA-3’</td>
</tr>
<tr>
<td>LTRIII</td>
<td>5’-AGTAGTGCTGTGCCGCTCCTGTCTG-3’</td>
</tr>
<tr>
<td>LCI</td>
<td>5’-GACCCGGGAGATCTGAATTCAGAGCAGTTAG-3’</td>
</tr>
<tr>
<td>LCII</td>
<td>5’-GACCCGGGAGATCTGAATTCAGAGCAGTTAG-3’</td>
</tr>
</tbody>
</table>

<sup>a</sup>LM-PCR, ligation-mediated PCR

<sup>b</sup>All oligonucleotides were purchased from Sigma-Aldrich.
primers LTRII and LCI (Table 3.4). Cycles consisted of a denaturing step (95°C for 5 minutes), followed by 35 cycles of template denaturing (95°C for 60 seconds), primer annealing (60°C for 45 seconds) and extension (72°C for 90 seconds), followed by a final extension step (72°C for 10 minutes). Two microlitres of this initial PCR product were used as template for a nested PCR, which utilised identical conditions to the initial PCR, except that the primers LTRIII and LCII were used in place of primers LTRII and LCI (Table 3.4).

Nested PCR products were visualised by agarose gel electrophoresis (Section 2.3.3.2). Bands were gel purified and Sanger-sequenced directly using primer LCII if fragments were greater than 120 bp (Section 2.3.7). Smaller nested PCR products were subcloned into the TOPO-TA vector (Section 2.3.6) for Sanger-sequencing (Section 2.3.7).

3.3.4.3 Microarray analysis of gene expression

RNA was extracted from pelleted lymphoma cells (Sections 3.3.3 and 2.3.2.5) and sent to the Ramaciotti Centre for Gene Function Analysis (Randwick, Australia). Microarray analysis was performed using an Affymetrix mouse gene array 2.1 for mouse (Affymetrix, Santa Clara, USA). To serve as control samples, RNA was extracted from adult wild-type thymocytes or splenic T cells selected using a CD90.2 positive selection kit (StemCell Technologies). The immunophenotype of the lymphoma sample was used to select the most appropriate control sample during subsequent data analysis. The thymocyte sample was used as the control for lymphomas with an immature T cell phenotype, while the splenic T cell sample was used as the control for lymphomas with a mature T cell phenotype.
Chapter 3: Evaluating the safety and efficacy of lentiviral vectors in vivo

Data for each sample were normalised using Expression Console 1.2.1.20 (Affymetrix) according to the Robust Multiarray Averaging (RMA) method, which applies the quantile normalisation approach (Irizarry et al. 2003). Genbank Accession Numbers for genes within a 500 kb window in either direction of the lentiviral vector integration sites that were identified in each lymphoma sample were obtained using the Table Browser function of UCSC Genome Bioinformatics <http://www.genome.ucsc.edu/cgi-bin/hgTables?command=start>. The ratio of signal intensity for the lymphoma sample relative to the control sample was calculated for each probe that corresponded to each of the Genbank Accession Numbers of interest. Lymphoid oncogenes were annotated using the allOnco Cancer Gene List <http://www.bushmanlab.org/links/genelists>, which is a compilation of all cancer-related genes from several databases.

3.3.5 Preparation of vector-genomic junction fragment libraries representing lentiviral vector integration sites for next generation sequencing

An LM-PCR-based method for amplifying LTR-genomic junction fragments for subsequent next generation sequencing has been described (Figure 3.2; Ciuffi et al. 2009). This method has been modified to improve the efficiency and specificity of linker cassette ligation to Tsp509I-digested genomic DNA and adapted for sequencing of nested PCR products using the Illumina NGS platform.
Figure 3.2: Methodology for preparing libraries of LTR-genomic DNA fragments for integration site analysis. Genomic DNA containing integrated lentiviral provirus was digested using Tsp509I and the sticky overhangs were partially filled using dATP. A linker cassette was then ligated to the TT overhangs. Fragments containing the 5’LTR and proviral sequence downstream were cleaved using SacI to prevent amplification of these fragments that did not contain genomic DNA. Vector-genomic fragments representing the junction between the 3’LTR and genomic DNA were then specifically amplified using LTR-specific primers and a linker-specific primers in an initial PCR and then a subsequent nested PCR. LTR, long terminal repeat.
3.3.5.1 Extraction and digestion of genomic DNA from EF1α.γc-treated and PGK.γc-treated mice

*IL2Rγ^-/-Rag2^-/-* mice treated with the EF1α.γc and PGK.γc lentiviral vectors (Section 3.4.1) were bled at 12 weeks post-transplantation for subsequent extraction of genomic DNA (Section 2.3.2.4).

Genomic DNA (0.7-7.0 μg) was digested with *Tsp509I* for 6 hours, and genomic DNA fragments were organically extracted (Section 2.3.3.5) and precipitated with ethanol (Section 2.3.3.6). The 4 base 5'-AATT overhangs produced by the *Tsp509I* digest were partially filled in using the 3'→5' exo- Klenow Fragment (New England Biolabs) and dATP (1 mM).

3.3.5.2 Preparation and ligation of a linker cassette to *Tsp509I*-digested genomic DNA

A *Tsp509I*-compatible linker cassette was prepared by annealing the oligonucleotides HincII and Tsp509I-filled (Table 3.5) at final concentrations of 40 μM each in 10 mM Tris pH 8.0, 0.1 mM EDTA.

The following thermocycler conditions were employed to anneal the oligonucleotides using an Eppendorf Mastercycler Gradient PCR machine: 2 minutes at 92°C, followed by a temperature decrease in increments of 0.1°C every 4 seconds to 82°C, every 5 seconds to 72°C, every 8 seconds to 62°C, every 10 seconds to 52°C, every 12 seconds to 42°C and every 15 seconds to 12°C. The linker cassette was ligated to the *Tsp509I*-digested genomic fragments at a 40-fold molar excess overnight at 16°C using T4 DNA ligase (New England Biolabs; Section 2.3.5.2). Following heat inactivation of T4 ligase, DNA in the ligation reaction was precipitated with ethanol to
Table 3.5: Oligonucleotide sequences used for preparation of vector-genomic junction fragment libraries

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HincII</td>
<td>5'-GTAATACGACTCACTATAGGGCACGCGTGTCGACGACGCCCAGGC-3'</td>
</tr>
<tr>
<td>Tsp509I-filled</td>
<td>5'[Phos]-TTGCAGCCCG-3'[AmC7]</td>
</tr>
<tr>
<td>Lenti-LTR-outer</td>
<td>5'-TGCTTCTCAAGTAGTGTCGACGCCCAGGC-3'</td>
</tr>
<tr>
<td>LNT1</td>
<td>5'-CAACGAGAAGACCGACAGATCTTTCAGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>L1</td>
<td>5'-GACCTCAGAATAGGCAGGCG-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-1</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-2</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-3</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-4</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-5</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-6</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-7</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-8</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-9</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-10</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-11</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
<tr>
<td>Lenti-LTR-nested-12</td>
<td>5'-ATGATACGACACCACAGATCTACACTTTCTTCACTACGAGCATCGCTTCTCCAGTGTCGACGACGCCCCAGGC-3'</td>
</tr>
</tbody>
</table>

*a*All oligonucleotides were purchased from Sigma-Aldrich.

*b*Illumina adaptor sequences are underlined and barcode indexes used for multiplexing of samples are italicised.
remove salts (Section 2.3.3.6). The DNA was digested with SacI to cleave ‘internal control’ fragments, generated by ligation of the linker cassette to the first Tsp509I-end downstream of the vector 5’LTR. These contained vector sequence between the LTR and linker cassette sequences, rather than genomic sequence. An organic extraction (Section 2.3.3.4) and ethanol precipitation (Section 2.3.3.6) were performed after this second digest to remove the SacI enzyme and salts.

3.3.5.3 Initial PCR amplification of vector-genomic DNA

The linkeredin 3’-LTR-genomic fragments were used as template for PCR amplification (LM-PCRI) using the lentiviral LTR-binding primer Lenti-LTR-outer and the linker-binding primer L1 (Table 3.5). A 50 μL reaction used approximately 500 ng genomic DNA, 2.5 U HotStarTaq Plus DNA polymerase (Qiagen), reaction buffer, MgCl₂ (2 mM), the 4 dNTPs (0.2 mM) and primers lenti-LTR-outer and L1 (0.3 μM). Cycling conditions followed a touchdown PCR format and consisted of an initial denaturing step at 95°C for 5 minutes, followed by 7 cycles of template denaturing (94°C for 30 seconds) and primer annealing and extension (72°C for 60 seconds), then 37 cycles of template denaturing (94°C for 30 seconds) and primer annealing and extension (68°C for 60 seconds), then a final extension step (68°C for 3 minutes). PCR products of 120–400 bp were excised from a 2.5% TAE agarose gel and purified (Sections 2.3.3.2 and 2.3.3.4).

3.3.5.4 Nested PCR amplification of vector-genomic DNA

Five microlitres of 1:10 and 1:100 dilutions of purified products from LM-PCRI were used as template for a nested PCR (nPCR). The nPCR used the lentiviral LTR-binding primer Lenti-LTR-nested and the linker-binding primer LNT1 (Table 3.5). Twelve
different barcoded LTR-nested primers were used to enable multiplexing of samples that originated from 12 different mice. Reagent and cycling conditions for the nPCR were identical to those for LM-PCRI, except a 25 μL final volume was used. nPCR products that were 175–500 bp were excised from a 2.5% TAE agarose gel and purified (Sections 2.3.3.2 and 2.3.3.4). Purified nPCR products derived from 2 representative mouse samples were subcloned using the TOPO-TA Cloning Kit (Section 2.3.6) and Sanger-sequenced (Section 2.3.7).

3.3.6 Next generation sequencing of vector-genomic DNA libraries and data analysis

The size-selected nPCR products (Section 3.3.5.4) represented 3'LTR-genomic junction fragment libraries for 12 different mice. Each library was mixed in equimolar ratio and sequenced using an Illumina HiSeq2000 instrument (Illumina Inc., San Diego, USA) at the Beijing Genomics Institute, Hong Kong, China.

Analysis of the sequence data was performed by Dr Gagan Garg and Mr Mohammad Islam and proceeded as follows, with preliminary filtering and trimming of sequence reads before those reads were mapped to the mouse genome (Figure 3.3). Sequence reads were separated for each of the 12 samples according to the 4 nucleotide (nt) barcode indices, which were then trimmed. Sequence reads containing the expected LTR sequence at nt positions 1-27, “CCCTTTTAGTCAGTGTGGAAAATCTCT”, were retained, allowing up to 3 mismatches and/or base omissions. Those 27 nt were then trimmed. Sequence reads were next filtered for reads containing the final 4 nt of the LTR, “AGCA”, the presence of which increased confidence that the amplicons represented by the sequence reads were derived from genuine vector-genomic fragments, rather than non-specific amplification. Those 4 nt were also trimmed.
Chapter 3: Evaluating the safety and efficacy of lentiviral vectors in vivo

Figure 3.3: Structure of integration site libraries and bioinformatic workflow for analysis of sequence data. (A) Integration site libraries contained adaptor sequences for the Illumina sequencing platform, then a 4 nt barcode index used to identify samples, followed by 31 nt of the lentiviral LTR sequence, then the genomic sequence of unknown length including the Tsp509I recognition site “AATT”, then the linker cassette sequence. Since 100 base read lengths were used, sequence reads contained either the whole, a portion or none of the linker cassette sequence. (B) Raw sequence reads from the Illumina HiSeq2000 were separated using the barcode index, then filtered for the expected LTR sequence in the first 27 nt (allowing 3 mismatches), which were trimmed. Reads were then filtered for the last 4 nt of the LTR, which were also trimmed, before whole or part of the linker cassette sequence was trimmed using the partial sequence “AATTGCAG”. Any remaining reads shorter than 18 nt were discarded, as these were unlikely to map unequivocally to the genome. Processed reads were mapped to the mouse genome (Assembly mm10) using Bowtie2, and end-to-end alignments with 0 mismatches tolerated. nt, nucleotide; LTR, long terminal repeat.
Finally, if any portion of the linker cassette was present it was trimmed by searching for “AATTGCAG” and trimming that sequence as well as any following sequence.

The remaining portion of the sequence reads contained genomic sequence, with the first nucleotide representing the integration site. Any reads shorter than 18 nt in length were discarded to increase the likelihood of mapping reads uniquely to the genome. Filtered reads were mapped to the mouse genome (Assembly mm10) using the alignment program Bowtie2 (Langmead & Salzberg 2012). End-to-end alignments were performed and zero mismatches were tolerated. The number of unique integration sites were counted, and a correction for the different amounts of input genomic DNA was applied.

3.3.7 Statistical analysis

Statistical comparisons of lymphocyte reconstitution between mice treated with the EF1α.γc and PGK.γc vectors or with wild-type Sca1⁺ BM cells were performed using an unpaired 2-tailed Student t-test, analysis of variance of groups (ANOVA), the Mann-Whitney test and the Kruskal-Wallis test (GraphPad Prism Version 5). The total numbers of unique integration sites in mice treated with the EF1α.γc and PGK.γc vectors, corrected for the amount of input lymphocyte DNA, were also compared using an unpaired 2-tailed Student t-test. Survival analysis was conducted using the Kaplan-Meier method and survival curves were compared using a 2-tailed log-rank (Mantel-Cox) test (GraphPad Prism Version 5). For comparisons of the incidence of lymphomagenesis, the 2-tailed Fisher’s exact test was used and for comparisons of the onset of lymphomagenesis, the unpaired 2-tailed Student t-test was used.
3.4 RESULTS

3.4.1 Setting up of cohorts of \textit{IL2R}\textsubscript{\gamma}-deficient mice treated with EF1\textalpha.\gamma\textsubscript{c} and PGK.\gamma\textsubscript{c} lentiviral vectors

Lentiviral vectors with either the EF1\textalpha or PGK promoter-enhancer elements controlling \gamma\textsubscript{c} expression, EF1\textalpha.\gamma\textsubscript{c} and PGK.\gamma\textsubscript{c}, have been described (Figure 3.4A; Smyth \textit{et al}. 2007; Ginn \textit{et al}. 2010). The EF1\textalpha promoter-enhancer was 5.5-fold more transcriptionally active than the PGK promoter-enhancer in human lymphoid ED-7R cells when 17% of cells were transduced (Figure 3.4B). \textit{IL2R}\textsubscript{\gamma}\textsuperscript{-/-} \textit{Rag2}\textsuperscript{-/-} mice were transplanted with \textit{IL2R}\textsubscript{\gamma}\textsuperscript{-/-} Sca1\textsuperscript{+} BM cells, which had been transduced with either the EF1\textalpha.\gamma\textsubscript{c} vector (n = 27) or the PGK.\gamma\textsubscript{c} vector (n = 26; Figure 3.1). \textit{IL2R}\textsubscript{\gamma}\textsuperscript{-/-} \textit{Rag2}\textsuperscript{-/-} mice were also transplanted with an equivalent dose of wild-type Sca1\textsuperscript{+} BM cells (n = 27).

3.4.2 Reconstitution of immunity in \textit{IL2R}\textsubscript{\gamma}-deficient mice treated with EF1\textalpha.\gamma\textsubscript{c} and PGK.\gamma\textsubscript{c} lentiviral vectors

Levels of lymphocyte reconstitution, including T cell, B cell and NK cell subsets, were measured in \textit{IL2R}\textsubscript{\gamma}\textsuperscript{-/-} \textit{Rag2}\textsuperscript{-/-} mice treated with the EF1\textalpha.\gamma\textsubscript{c} and PGK.\gamma\textsubscript{c} vectors at 6 weeks and 16 weeks after transplantation. At 6 weeks post-transplantation, overall levels of lymphocyte reconstitution were lower in mice treated with the PGK.\gamma\textsubscript{c} vector compared to the EF1\textalpha.\gamma\textsubscript{c} vector (Figure 3.5). In particular, B cell counts were 16.2-fold lower in PGK.\gamma\textsubscript{c}-treated mice than they were in EF1\textalpha.\gamma\textsubscript{c}-treated mice. T cell counts were 3.9-fold lower in PGK.\gamma\textsubscript{c}-treated mice. NK cell reconstitution was poor in both EF1\textalpha.\gamma\textsubscript{c}-treated and PGK.\gamma\textsubscript{c}-treated mice. Mice treated with a wild-type transplant displayed the highest counts of all lymphocyte subsets.
Figure 3.4: Schematic diagram of self-inactivating lentiviral vectors EF1α.γc and PGK.γc, and relative strengths of the EF1α and PGK promoter-enhancer elements used. (A) Lentiviral vectors in which expression of the IL-2 receptor common γ chain (γc) is under the transcriptional control of the elongation factor-1-α (EF1α) or phosphoglycerate kinase (PGK) promoter-enhancers. HIV-1 sequences retained in the vectors include GAG, the Rev response element (RRE), the central polypurine tract (cPPT), splice donor (SD) and acceptor (SA) sites, and packaging and dimerisation signal (Ψ). (B) Human lymphoid ED-7R cells were transduced using lentiviral vectors in which enhanced green fluorescent protein (EGFP) expression is controlled by either of the EF1α or PGK promoter-enhancers. When approximately 17% of cells were transduced, the mean fluorescence intensities for the EF1α and PGK promoter-enhancers were 17,500 and 3,200, respectively, indicating the EF1α promoter-enhancer is 5.5-fold more active than the PGK promoter-enhancer in ED-7R cells.
Figure 3.5: Levels of reconstituted lymphocyte subsets at 6 weeks post-transplantation in γc-deficient mice treated with EF1α.γc (n = 27) and PGK.γc (n = 26) lentiviral vectors, or with a wild-type transplant (n = 27). Counts of total reconstituted (A) CD45⁺CD2⁺ lymphocytes, (B) CD3⁺NK1.1⁻ T cells, (C) B220⁺CD3⁻ B cells, and (D) NK1.1⁺CD3⁻ NK cells. T cell counts for mice that subsequently developed lymphoma are coloured red. Overall, reconstitution was lower for the PGK.γc vector compared to the EF1α.γc vector: 8.9-fold lower for total lymphocytes, 3.9-fold lower for T cells, 16.2-fold lower for B cells, and 6.1-fold lower for NK cells. Mean reconstitution levels were lower compared to wild-type transplant in mice treated with EF1α.γc and PGK.γc for total lymphocytes (2.4-fold lower and 21.0-fold lower, respectively), B cells (2.4-fold lower and 38.4-fold lower, respectively) and NK cells (15.3-fold lower and 91.7-fold lower, respectively). T cell counts were lower than wild-type in mice treated with PGK.γc (5.6-fold), but not in mice treated with EF1α.γc. ANOVA analysis of mean reconstitution levels indicated significant differences for total lymphocyte (F (2,77) = 44.7, p < 0.0001), T cell (F (2,76) = 19.0, p < 0.0001), B cell (F (2,77) = 35.4, p < 0.0001), and NK cell (F (2,76) = 57.4, p < 0.0001) subset counts. Error bars represent standard errors of the mean. **, 0.001 < p < 0.01; ***, 0.0001 < p < 0.001; **** p < 0.0001, calculated using the Student t test. ANOVA, analysis of variance of between groups; EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; NK, natural killer; ns, not significant; PGK, phosphoglycerate kinase.
By 16 weeks post-transplantation, peripheral blood lymphocytes in the PGK.γc cohort (n = 28) had increased by 2-fold, but were still 5.5-fold lower than lymphocyte counts in the EF1α.γc cohort (Figure 3.6). Lymphocyte reconstitution in EF1α.γc-treated mice (n = 28) was half of that achieved in mice that received a wild-type transplant (n = 27). Overall, counts of each of the lymphocyte subsets were more varied for the cohorts treated by gene therapy compared to the mice that received a wild-type transplant. T cell counts expanded by 2.6-fold and 3.4-fold between the 6 week and 16 week time-points for the EF1α.γc and PGK.γc cohorts, respectively. Absolute T cell counts of PGK.γc-treated mice were approximately one-third of T cell counts in EF1α.γc-treated mice, which were 1.8-fold lower than T cell counts in wild-type-transplanted mice.

B cell reconstitution was impaired in mice treated with the PGK.γc vector compared to the mice treated with the EF1α.γc vector, with 18 of the PGK.γc-treated mice having few or no reconstituted B cells. B cell reconstitution was also reduced for the EF1α.γc vector relative to the wild-type-transplanted mice, with counts in EF1α.γc-treated mice at less than half of those of wild-type-transplanted mice. Levels of NK cell reconstitution in both EF1α.γc-treated and PGK.γc-treated mice were poor in comparison to mice treated with a wild-type transplant.

3.4.3 Skewed T cell to B cell ratio and CD4⁺ T cell to CD8⁺ T cell ratio in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

To explore the extent to which B cell reconstitution was impaired in vector-treated cohorts, the ratios of T cells to B cells were calculated for mice treated with the EF1α.γc and PGK.γc vectors, as well as the wild-type-transplanted mice. The PGK.γc vector had an approximately 7.8-fold higher mean T cell to B cell ratio compared to the
Figure 3.6: Levels of reconstituted lymphocyte subsets at 16 weeks post-transplantation in γc-deficient mice treated with EF1α.γc (n = 28) and PGK.γc (n = 28) lentiviral vectors, or with a wild-type transplant (n = 27). Counts of total reconstituted (A) CD45+CD25+ lymphocytes, (B) CD3+NK1.1− T cells, (C) CD3+CD8+CD4+ T cells, (D) CD3+CD4+CD8+ T cells, (E) B220+CD3+ B cells, and (F) NK1.1+CD3− NK cells. Mean reconstituted lymphocyte counts were lower for the PGK.γc vector compared to the EF1α.γc vector: 5.5-fold lower for total lymphocytes, 2.9-fold lower for T cells, 23.2-fold lower for B cells, and 8.5-fold lower for NK cells. Mean reconstitution levels were lower compared to wild-type transplant in mice treated with EF1α.γc and PGK.γc for total lymphocytes (1.9-fold and 10.4-fold, respectively), T cells (1.8-fold and 1.7-fold, respectively), B cells (2.4-fold and 55.5-fold, respectively), and NK cells (19.7-fold and 161.5-fold, respectively). Mean CD8+ T cell counts did not differ between EF1α.γc-treated mice and wild-type-treated mice. CD4+ T cell counts did not differ between PGK.γc-treated mice and wild-type-treated mice, however, CD8+ T cell counts were 3.3-fold lower in PGK.γc-treated mice. ANOVA analysis of mean reconstitution levels indicated significant differences for total lymphocyte (F (2,80) = 61.3, p < 0.0001), T cell (F (2,88) = 9.2, p = 0.0002), B cell (F (2,83) = 27.9, p < 0.0001), and NK cell (F (2,83) = 69.4, p < 0.0001) subset counts. Error bars represent standard errors of the mean. *, p < 0.05; **, 0.001 < p < 0.01; ***, 0.0001 < p < 0.001; **** p < 0.0001, calculated using the Student t test. ANOVA, analysis of variance of between groups; EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; NK, natural killer; ns, not significant; PGK, phosphoglycerate kinase.
EF1α.γc vector (p = 0.0054; Figure 3.7A). The T cell to B cell ratios had a wider spread for the PGK.γc vector compared to the EF1α.γc vector. The ratio of T cells to B cells was 16.3-fold higher in EF1α.γc-treated mice compared to wild-type-transplanted mice (p = 0.045).

Ratios of CD4+ T cells and CD8+ T cells were also calculated for all mice. Reconstituted T cells in mice treated with PGK.γc were skewed towards the CD4+ subset. For the PGK.γc vector, the mean ratio of CD4+ T cells to CD8+ T cells was 4-fold higher than that of the EF1α.γc vector (p = 0.0026) and 5-fold higher than that of the wild-type transplant (p = 0.0004; Figure 3.7B). The mean ratio of CD4+ T cells to CD8+ T cells did not differ between the EF1α.γc-treated mice and the wild-type-transplanted mice.

3.4.4 Long-term safety monitoring of IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

Cohorts of mice treated with the EF1α.γc (n = 23) and PGK.γc (n = 14) vectors, or transplanted with wild-type Sca1+ BM cells (n = 27), were maintained for over 15 months to monitor for long-term safety. Lymphomas developed in 6 mice from the EF1α.γc cohort and in 3 mice from the PGK.γc cohort, but in none of the mice transplanted with wild-type Sca1+ BM cells (Figure 3.8). The onset of lymphomagenesis in mice treated with EF1α.γc and PGK.γc ranged from 6.5–13.6 months post-transplantation (Table 3.6).
Figure 3.7: Ratios of CD3^+ T cells to B220^+ B cells, and of CD4^+ T cells to CD8^+ T cells in mice treated with EF1α.γc, PGK.γc and wild-type Sca1^+ bone marrow cells. (A) The mean ratio of T cell counts to B cell counts for the EF1α.γc vector, at 8.48 ± 3.73, was lower than that of the PGK.γc vector, at 66.50 ± 19.65. Both ratios were higher than the ratio of T cell counts to B cell counts in recipients of wild-type cells, at 0.21 ± 0.02. ANOVA analysis indicated significantly different mean ratios of T cell counts to B cell counts (p < 0.0001, Kruskall-Wallis test). (B) The mean ratio of CD4^+ to CD8^+ T cells was 1.08 ± 0.29 for the EF1α.γc vector and 4.32 ± 1.00 for the PGK.γc vector. The mean ratio of CD4^+ T cells to CD8^+ T cells for the PGK.γc vector was lower than that of mice receiving a wild-type transplant, at 0.83 ± 0.07. There was no significant difference between the EF1α.γc and wild-type cohorts in the mean ratio of CD4^+ T cells to CD8^+ T cells. ANOVA analysis indicated significantly different mean ratios of CD4^+ to CD8^+ T cells (p < 0.0001, Kruskall-Wallis test). Error bars represent standard errors of the mean. *, p < 0.05; **, 0.001 < p < 0.01; ***, 0.0001 < p < 0.001; ****, p < 0.0001, calculated using the Mann-Whitney test. ANOVA, analysis of variance of between groups EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; ns, not significant; PGK, phophoglycerate kinase.
Figure 3.8: Proportion of lymphoma-free survival in mice treated with EF1α.γc or PGK.γc lentiviral vectors, or with wild-type Sca1+ bone marrow cells. Survival was not significantly different between the EF1α.γc and PGK.γc cohorts (Hazard ratio = 1.27 [0.33–4.86], p = 0.724). Survival was significantly lower for the EF1α.γc vector compared to the wild-type transplant (Hazard ratio = 9.74 [1.94–48.88], p = 0.0057). Survival was also significantly lower for the PGK.γc vector compared to the wild-type transplant (Hazard ratio = 20.19 [1.83–223.0], p = 0.0142). Confidence intervals (95%) for hazard ratios are given in square brackets. EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; PGK, phosphoglycerate kinase.
### Table 3.6: Macroscopic and cellular characteristics of lymphomas arising in \( IL2R_{\gamma} \)-deficient mice treated with \( EF1_{\alpha},\gamma c \) and \( PGK,\gamma c \) lentiviral vectors

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Vector</th>
<th>Latency (months post-transplant)</th>
<th>White cell count (per μL)</th>
<th>White cell count fold-expansion(^a)</th>
<th>Immunophenotype(^b)</th>
<th>Spleen mass (g)</th>
<th>Thymus mass (g)</th>
<th>Liver mass (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m196</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>7.5</td>
<td>21,480</td>
<td>15.30</td>
<td>( CD3^{low}CD4^{+}CD8^{+} )</td>
<td>0.22</td>
<td>0.29</td>
<td>1.71</td>
</tr>
<tr>
<td>m234</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>6.5</td>
<td>18,720</td>
<td>7.15</td>
<td>( CD3^{low}CD4^{(low)}CD8^{+} )</td>
<td>0.44</td>
<td>0.13</td>
<td>6.09</td>
</tr>
<tr>
<td>m115</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>13.5</td>
<td>27,040</td>
<td>12.07</td>
<td>( CD3^{low}CD4^{+}CD8^{+}B220^{+} )</td>
<td>0.43</td>
<td>0.06</td>
<td>2.26</td>
</tr>
<tr>
<td>m236</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>9.0</td>
<td>7,000</td>
<td>3.67</td>
<td>( CD3^{low}CD4^{+}CD8^{+} )</td>
<td>0.34</td>
<td>0.17</td>
<td>2.36</td>
</tr>
<tr>
<td>m308</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>8.5</td>
<td>2,150</td>
<td>0.67</td>
<td>( CD3^{+}CD4^{+}CD8^{+} )</td>
<td>0.31</td>
<td>0.15</td>
<td>1.64</td>
</tr>
<tr>
<td>m54</td>
<td>( EF1_{\alpha},\gamma c )</td>
<td>10.4</td>
<td>6,100</td>
<td>1.04</td>
<td>( CD3^{low}CD4^{low}CD8^{+} )</td>
<td>0.34</td>
<td>0.36</td>
<td>2.93</td>
</tr>
<tr>
<td>m241</td>
<td>( PGK,\gamma c )</td>
<td>7.8</td>
<td>6,920</td>
<td>4.27</td>
<td>( CD3^{low}CD4^{+}CD8^{+} )</td>
<td>0.75</td>
<td>0.06</td>
<td>1.56</td>
</tr>
<tr>
<td>m24</td>
<td>( PGK,\gamma c )</td>
<td>11.1</td>
<td>107,333</td>
<td>17.28</td>
<td>( CD3^{+}CD4^{+}CD8^{+} )</td>
<td>0.10</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>m29</td>
<td>( PGK,\gamma c )</td>
<td>9.1</td>
<td>6,275</td>
<td>3.70</td>
<td>( CD3^{low}CD4^{+}CD8^{+} )</td>
<td>0.07</td>
<td>0.37</td>
<td>0.69</td>
</tr>
</tbody>
</table>

\(^a\)White cell count fold-expansion is calculated relative to the white cell count measured at 16 weeks post-transplant.

\(^b\)With the exception of m115, all cases of lymphoma had a B220\(^-\) immunophenotype.
Survival was not significantly different between the EF1α.γc and PGK.γc cohorts (Hazard ratio = 1.27, p = 0.724), and nor were the incidence (p = 1.000) or onset (p = 0.7612) of lymphomagenesis. When compared to the cohort transplanted with wild-type Sca1⁺ BM, on the other hand, survival was significantly lower for the cohorts treated with the EF1α.γc and PGK.γc vectors (Hazard ratio = 9.74 and p = 0.0057, Hazard ratio = 20.19 and p = 0.0142, respectively). Similarly, the incidence of lymphomagenesis was significantly higher for the cohorts treated with EF1α.γc and PGK.γc compared to the cohort transplanted with wild-type Sca1⁺ BM (p = 0.0072 and p = 0.0368, respectively). Therefore, although both the EF1α.γc and PGK.γc vectors were associated with a risk of lymphomagenesis, there was no difference in the frequency of this risk between the 2 vectors.

3.4.4.1 Macroscopic and cellular characteristics of lymphomas arising in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

In the mice that developed lymphoma, white cell counts had expanded by 3.7-fold to 17.3-fold compared to the white cell count measured at 16 weeks post-transplant, with the exception of 2 cases of lymphoma in the EF1α.γc cohort (Table 3.6). All cases had a predominantly T cell phenotype with varying degrees of immaturity, as assessed by double positivity for CD4 and CD8. Thymuses were detected in all of the mice that developed lymphoma, even though these mice were over 12 months in age. Thymuses are usually difficult to see macroscopically in similarly-aged mice treated with wild-type or IL2Rγ-modified Sca1⁺ BM due to thymic involution. All mice had enlarged spleens and livers, except for 2 mice in the PGK.γc cohort. Tumour infiltration and disruption of organ architecture were observed in the spleen, liver, lungs and kidneys of all mice, and also in thymus where sufficient tissue was available (Figure 3.9).
Figure 3.9: Lymphoma infiltration of tissue. For mice treated with EF1α.γc or PGK.γc lentiviral vectors that developed lymphomas, tissue from spleen, liver, kidney, lung and thymus (where available) was stained with haematoxylin and eosin. Control tissue was obtained from a mouse treated with EF1α.γc, which developed an external tumour that was unrelated to gene therapy, at 9.2 months post-transplantation. Scale bar is 100 μm. Ctrl, age-matched control; EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; PGK, phosphoglycerate kinase.
3.4.4.2 Analyses of vector copy numbers, vector integration sites and gene dysregulation in lymphomas

For each of the lymphomas, average VCN ranged from 0.73–1.71, and 1–2 vector integration sites were detected by LM-PCR (Table 3.7). Vector integrations were identified in both forward and reverse orientations, across a range of different chromosomes, and occurred both within and between genes of diverse functions.

Microarray analysis of RNA extracted from each lymphoma sample did not uncover evidence of gene dysregulation in the vicinity of vector integration sites. For all of the lymphoma samples, a total of 109 genes were identified within 500 kb upstream or downstream of vector integration sites. The relative intensities of the probe sets representing those 109 genes ranged from 0.55–2.27 (Figure 3.10), and were at least 1 log-fold lower than the insertionally dysregulated expression of LMO2 and CCND2 in the French and British SCID-X1 trials (Hacein-Bey-Abina et al. 2008; Howe et al. 2008). Although 11 of those 109 genes were classified as lymphoid oncogenes, the mean relative expression of these lymphoid oncogenes was 1.09. Over-expression of IL2Rγ was not detected in any of the samples. The mean relative intensity of the probe corresponding to IL2Rγ was 0.56 ± 0.03 (standard deviation).

Of the 109 genes within 1 megabase (Mb) of the vector integration sites, those with a relative expression that was at least 1 standard deviation above the mean were selected for further comparisons against each of the other tumour samples. For each of those 10 genes, similar expression levels were detected in at least 2 other tumour samples, which frequently displayed a similar immunophenotype and contained vector integrations at different genomic loci, often in a separate chromosome. This implied that the dysregulation of those genes relative to the non-malignant control samples was
### Table 3.7: Characteristics of vector integrations in lymphomas arising in $IL2R_{\gamma}$-deficient mice treated with $EF1_{\alpha,\gamma c}$ and $PGK_{\gamma c}$ lentiviral vectors

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Vector</th>
<th>Vector copy number (qPCR)</th>
<th>Vector integrations detected by LM-PCR</th>
<th>Vector orientation</th>
<th>Chromosome of integration site</th>
<th>Intergenic or intragenic integration site</th>
<th>Details of genes nearest to the integration site</th>
</tr>
</thead>
<tbody>
<tr>
<td>m196</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>1.11</td>
<td>1</td>
<td>Forward</td>
<td>X</td>
<td>Intergenic</td>
<td>Hypothetical protein LOC100504400 is 1,160.0kb upstream; glutamate receptor 3 precursor is 377.9kb downstream</td>
</tr>
<tr>
<td>m234</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>0.80</td>
<td>1</td>
<td>Reverse</td>
<td>3</td>
<td>Intragenic</td>
<td>Calcium/calmodulin-dependent protein kinase II delta $CAMK2D$ intron 3</td>
</tr>
<tr>
<td>m115</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>1.53</td>
<td>1</td>
<td>Forward</td>
<td>X</td>
<td>Intergenic</td>
<td>Predicted gene 2799 is 22.3 kb upstream; germ cell-less protein-like 1-like is 257.7 kb downstream</td>
</tr>
<tr>
<td>m236</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>1.01</td>
<td>1</td>
<td>Reverse</td>
<td>3</td>
<td>Intragenic</td>
<td>B cell scaffold protein with ankyrin repeats $Bank1$ intron 6</td>
</tr>
<tr>
<td>m308</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>2.20</td>
<td>1</td>
<td>Reverse</td>
<td>8</td>
<td>Intergenic</td>
<td>Vascular endothelial growth factor C precursor 165.9 kb upstream; signal peptidase complex subunit 3171.2 kb downstream</td>
</tr>
<tr>
<td>m54</td>
<td>$EF1_{\alpha,\gamma c}$</td>
<td>1.71</td>
<td>2</td>
<td>Reverse</td>
<td>9</td>
<td>Intragenic</td>
<td>Transcription factor 12 tcf12 intron 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Forward</td>
<td>2</td>
<td>Intragenic</td>
<td>Amyloid beta A4 precursor protein-binding family B $Apbb1ip$ intron 4</td>
</tr>
<tr>
<td>m241</td>
<td>$PGK_{\gamma c}$</td>
<td>0.73</td>
<td>1</td>
<td>Forward</td>
<td>4</td>
<td>Intergenic</td>
<td>Coiled-coil domain-containing protein 21 is 2.5kb upstream; uromodulin-like is 15.7kb downstream</td>
</tr>
<tr>
<td>m29</td>
<td>$PGK_{\gamma c}$</td>
<td>0.80</td>
<td>1</td>
<td>Reverse</td>
<td>16</td>
<td>Intergenic</td>
<td>Intraflagellar transport protein 57 homolog is 30.5kb upstream; leukocyte surface antigen CD47 is 61.1kb downstream</td>
</tr>
<tr>
<td>m24</td>
<td>$PGK_{\gamma c}$</td>
<td>1.80</td>
<td>2</td>
<td>Forward</td>
<td>6</td>
<td>Intragenic</td>
<td>Transmembrane and coiled-coil domains protein 1 intron 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reverse</td>
<td>8</td>
<td>Intragenic</td>
<td>Vascular endothelial growth factor C precursor intron 1</td>
</tr>
</tbody>
</table>
Figure 3.10: Microarray analysis of the relative expression of genes within 1 Mb of lentiviral vector integration sites in lymphoma samples from mice treated with EF1α.γc or PGK.γc lentiviral vectors. For each sample, the intensities of microarray probe sets for genes within 1 Mb of vector integration sites were compared to the intensities of the same probe sets used to analyse non-malignant thymocytes or peripheral T cells. The relative intensity and location of each probe with respect to the vector integration is depicted. Vector integration sites are at position 0 on the x-axis and genomic coordinates of these sites are also annotated. Lymphoid oncogenes, annotated using the allOnco Cancer Gene List, are coloured red. chr, chromosome; EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; Mb, megabases; PGK, phosphoglycerate kinase.
more likely to have arisen from the tumour phenotype and was unrelated to vector integration. Therefore, there was no evidence of insertional dysregulation of any of those genes within a broad window, 1 Mb, of the lentiviral vector integration.

3.4.5 Selection of representative samples for preparation of vector-genomic DNA fragment libraries

To analyse the degree of clonal complexity in reconstituted lymphocytes from mice treated with the EF1α.γc and PGK.γc vectors, genomic DNA samples were selected from 6 mice from each of the EF1α.γc and PGK.γc cohorts. The samples selected represented mice whose T cell counts at 16 weeks post-transplantation were clustered around the mean T cell counts for each cohort and included 2 of the mice from each cohort that subsequently developed lymphoma. Integration site libraries were prepared using these selected samples (Figure 3.2).

3.4.6 Sequencing of integration site libraries and data processing

Integration site libraries representing the 12 mice selected for analysis were sequenced using an Illumina HiSeq2000 instrument, with a total of 36,584,663 100-base single end reads generated. Sequence reads were filtered based on the presence of the expected LTR sequence, which was trimmed before reads were mapped to the mouse genome (Figure 3.3). A total of 36,123,733 reads (98.7% of total reads generated) remained after filtering for reads containing the expected LTR sequence and 20,988,447 reads (57.4% of total reads generated) of at least 18 nt in length remained after trimming of whole or part of the linker cassette (Table 3.8). These reads, from which the LTR and linker cassette sequences were trimmed, were used for subsequent mapping to the genome.
Table 3.8: Numbers and proportions of sequence reads for each sample that were retained after each stage of data processing

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Vector</th>
<th>Barcode index</th>
<th>Total raw reads</th>
<th>Processed reads after trimming of LTR&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Processed reads after trimming of linker&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>m203</td>
<td>EF1α.γc</td>
<td>ATTG</td>
<td>1,974,679</td>
<td>1,954,927 (99.00%)</td>
<td>1,945,306 (98.51%)</td>
</tr>
<tr>
<td>m153</td>
<td>EF1α.γc</td>
<td>CGTG</td>
<td>2,468,923</td>
<td>2,440,286 (98.84%)</td>
<td>1,435,249 (58.13%)</td>
</tr>
<tr>
<td>m54</td>
<td>EF1α.γc</td>
<td>ACAT</td>
<td>2,675,104</td>
<td>2,643,086 (98.80%)</td>
<td>2,149,403 (80.35%)</td>
</tr>
<tr>
<td>m69</td>
<td>EF1α.γc</td>
<td>GCCT</td>
<td>3,590,981</td>
<td>3,540,516 (98.59%)</td>
<td>2,941,234 (81.91%)</td>
</tr>
<tr>
<td>m308</td>
<td>EF1α.γc</td>
<td>TGGT</td>
<td>3,645,769</td>
<td>3,600,204 (98.75%)</td>
<td>1,107,961 (30.39%)</td>
</tr>
<tr>
<td>m63</td>
<td>EF1α.γc</td>
<td>CACC</td>
<td>2,404,608</td>
<td>2,373,391 (98.70%)</td>
<td>1,663,151 (69.17%)</td>
</tr>
<tr>
<td>m399</td>
<td>PGK.γc</td>
<td>GATC</td>
<td>1,713,607</td>
<td>1,688,906 (98.56%)</td>
<td>709,700 (41.42%)</td>
</tr>
<tr>
<td>m38</td>
<td>PGK.γc</td>
<td>TCAA</td>
<td>2,173,072</td>
<td>2,146,577 (98.78%)</td>
<td>1,769,680 (81.44%)</td>
</tr>
<tr>
<td>m385</td>
<td>PGK.γc</td>
<td>CTGA</td>
<td>4,182,566</td>
<td>4,129,180 (98.72%)</td>
<td>2,788,002 (66.66%)</td>
</tr>
<tr>
<td>m29</td>
<td>PGK.γc</td>
<td>AGGC</td>
<td>4,841,941</td>
<td>4,783,417 (98.79%)</td>
<td>2,625,942 (54.23%)</td>
</tr>
<tr>
<td>m37</td>
<td>PGK.γc</td>
<td>GTAG</td>
<td>3,577,346</td>
<td>3,530,954 (98.70%)</td>
<td>1,569,627 (43.88%)</td>
</tr>
<tr>
<td>m24</td>
<td>PGK.γc</td>
<td>TACA</td>
<td>3,336,066</td>
<td>3,292,289 (98.69%)</td>
<td>283,192 (8.49%)</td>
</tr>
</tbody>
</table>

<sup>a</sup>The proportion in parentheses represents the number of reads with 27 nt of the LTR and the distal 4 final nt of LTR sequence as a percentage of total raw reads.

<sup>b</sup>The proportion in parentheses represents the number of reads with sequence as a percentage of total raw reads.

3.4.7 Mapping of unique integration sites in IL2Rγ-deficient mice treated with EF1α.γc and PGK.γc lentiviral vectors

Processed sequence reads were mapped to the mouse genome (Table 3.9). For 2 of the samples, m203 and m308, below 50% of trimmed sequence reads mapped to the genome. In m203, 74.2% of sequence reads were accounted for by 1 sequence that did not map to the mouse genome. Although it was possible that the unmapped reads may have contained internal lentiviral sequence amplified from the 5’LTR, analysis indicated this internal sequence was only present in approximately 1% of the unmapped reads.
### Table 3.9: Total mapped sequence reads and numbers of unique integration sites

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Vector</th>
<th>Total mapped sequence reads&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total unique integration sites per 500 ng lymphocyte DNA&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>m203</td>
<td>EF1α.γc</td>
<td>110,912 (5.70%)</td>
<td>69</td>
</tr>
<tr>
<td>m153</td>
<td>EF1α.γc</td>
<td>1,348,414 (93.95%)</td>
<td>1,285</td>
</tr>
<tr>
<td>m54</td>
<td>EF1α.γc</td>
<td>1595,898 (74.25%)</td>
<td>340</td>
</tr>
<tr>
<td>m69</td>
<td>EF1α.γc</td>
<td>2,669,334 (90.76%)</td>
<td>5,196</td>
</tr>
<tr>
<td>m308</td>
<td>EF1α.γc</td>
<td>353,011 (31.86%)</td>
<td>376</td>
</tr>
<tr>
<td>m63</td>
<td>EF1α.γc</td>
<td>1,408,752 (84.70%)</td>
<td>69</td>
</tr>
<tr>
<td>m399</td>
<td>PGK.γc</td>
<td>620,507 (87.43%)</td>
<td>707</td>
</tr>
<tr>
<td>m38&lt;sup&gt;c&lt;/sup&gt;</td>
<td>PGK.γc</td>
<td>1,714,996 (96.91%)</td>
<td>60,824</td>
</tr>
<tr>
<td>m385</td>
<td>PGK.γc</td>
<td>2,697,383 (96.75%)</td>
<td>2,418</td>
</tr>
<tr>
<td>m29</td>
<td>PGK.γc</td>
<td>2,436,111 (92.77%)</td>
<td>1,784</td>
</tr>
<tr>
<td>m37</td>
<td>PGK.γc</td>
<td>1,375,719 (87.65%)</td>
<td>6,901</td>
</tr>
<tr>
<td>m24</td>
<td>PGK.γc</td>
<td>200,730 (70.88%)</td>
<td>3,102</td>
</tr>
</tbody>
</table>

<sup>a</sup>The proportion in parentheses represents the number of reads that mapped to the mouse genome after LTR and linker cassette trimming, expressed as a percentage of total reads remaining after trimming of the LTR and linker cassette.

<sup>b</sup>The total number of unique integration sites identified within each sample was corrected for the different quantities of genomic DNA used to prepare vector-genomic libraries and for the proportion of that genomic DNA derived from the lymphocyte population.

<sup>c</sup>Although this value is an extreme outlier and is likely to skew the results, there is no biologically compelling reason to exclude it from the analysis.

The total number of unique integration sites detected within each vector-genomic library from each mouse sample was counted (Table 3.9). This was then corrected for the different quantities of input lymphocyte genomic DNA used for library preparation. Although there was no statistically significant difference in the mean number of unique integration sites identified, at $1,222 \pm 815.4$ (standard
deviation) for the EF1α.γc-treated mice and 12,622 ± 9,679 (standard deviation) for the PGK.γc-treated mice (p = 0.268), the variation within each cohort was so large that it renders any true differences between the 2 cohorts, should such differences exist, impossible to discern. Therefore, these results are inconclusive with regard to the comparison of PGK.γc versus EF1α.γc vector-associated reduction of lymphocyte complexity.

3.5 DISCUSSION OF CHAPTER 3 RESULTS

3.5.1 No significant difference in rates of lymphomagenesis between the EF1α.γc and PGK.γc lentiviral vectors

This chapter has shown that use of the PGK.γc lentiviral vector does result in reduced reconstitution of IL2Rγc/Rag2−/− mice compared to use of the EF1α.γc lentiviral vector, but not in a higher incidence of lymphomagenesis. Although the hypothesis underlying this chapter was that the transcriptionally less active PGK promoter-enhancer element would be associated with a higher rate of lymphomagenesis due to replicative stress, this hypothesis is not supported by the outcome of the long-term safety study described in this chapter. Intriguingly, the PGK.γc vector did have an effect on reconstitution levels and there are a few possibilities that could account for this outcome.

One possibility is that the reduced lymphocyte reconstitution associated with the transcriptionally less active PGK promoter-enhancer element implies that Sca1+ BM cells were not sufficiently stressed to cause a higher rate of lymphomagenesis. Consistent with this possibility is the observation that mice treated with PGK.γc had a smaller T cell compartment than mice treated with EF1α.γc. Had the T cell compartments been equivalent, it may have suggested that PGK.γc-transduced Sca1+
BM cells had proliferated more times in order to achieve a similarly-sized peripheral T cell compartment. The smaller-sized T cell compartment in PGK.γc-treated mice could imply that those Sca1+ BM cells were not under more pressure to replicate compared to Sca1+ BM cells in mice treated with the EF1α.γc vector. Similar levels of proliferation in Sca1+ BM cells in mice treated with either vector could account for the lack of difference in the incidence of lymphomagenesis attributable to replicative stress. Further experimentation could examine levels of markers for cellular proliferation, such as Ki67, in the bone marrow, thymocytes and peripheral lymphocytes of mice treated with either EF1α.γc or PGK.γc.

A second possibility could be that the effect of promoter-enhancer transcriptional activity on the rate of lymphomagenesis is small. If the effect size is small, it may be that the mouse model is insufficiently sensitive for making such measurements. The number of mice treated in this study may have been too small to measure the effect of promoter-enhancer transcriptional activity if the size of the effect is small. In order to measure a difference in effect between the proportions of mice that developed lymphoma in the 2 cohorts (21.4% for the PGK.γc-treated mice, 26.1% for the EF1α.γc-treated mice) with a statistical power (1 – β probability) of 0.5 and an α error probability of 0.05, it would be necessary to treat 492 mice in each cohort (G*Power 3, Faul et al. 2007). For a statistical power (1 – β probability) of 0.8 and an α error probability of 0.05, it would be necessary to treat 1,072 mice in each cohort.

In regard to the cohort sizes chosen for this study, the lack of knowledge about the effect of promoter-enhancer transcriptional activity on replicative stress limited the ability to estimate the effect size and hence calculate the number needed to treat. Therefore, the initial cohort sizes (n = 28 mice reconstituted at 16 weeks post-
transplantation) were as large as practicable, given the amount of time available. In light of the power analyses above, however, it is clear that these cohort sizes were too small to detect a difference in the incidence of lymphomagenesis between the EF1α.γc-treated and PGK.γc-treated mice. Notably, 14 of the PGK.γc-treated mice did not survive sufficiently long in order to be included in the long-term study. By contrast, 23 of the 28 engrafted EF1α.γc-treated mice survived sufficiently long to be included. Given that many of the PGK.γc-treated mice were partially reconstituted, it is possible that their compromised immune systems may have affected their survival. Thus, practical difficulties in maintaining large cohorts of immunocompromised mice impacted on how effectively the hypothesis about the effect of promoter-enhancer transcriptional activity on the incidence of lymphomagenesis could be addressed.

3.5.2 Inconclusive results of the integration site analysis for comparing clonal complexity in mice treated with the EF1α.γc and PGK.γc lentiviral vectors

A definitive conclusion cannot be drawn about the effect of the transcriptional activities of the EF1α and PGK promoter-enhancer elements on clonal complexity in the lymphocytes of mice treated with the EF1α.γc and PGK.γc vectors. The extent of within-cohort variation was large, at 69–5,196 and 707–60,824 unique integration sites for the EF1α.γc and PGK.γc vectors, respectively. This is likely to exceed any anticipated inter-cohort differences, given that the EF1α promoter-enhancer is only 5.5-fold more transcriptionally active than the PGK promoter-enhancer. Accordingly, it is not possible to meaningfully assess any inter-cohort differences. Therefore, these results are inconclusive with respect to the hypothesis that the transcriptionally less active PGK promoter-enhancer element results in reduced clonal complexity.
Technical and biological reasons may account for the variation in the final numbers of unique integration sites that were identified for the EF1α.γc and PGK.γc cohorts. The recovery of integration sites may have been biased by the known limitations associated with the use of restriction endonucleases and PCR (Section 1.5.4). The proportion of processed reads after LTR and linker cassette trimming was substantially reduced in some samples from each cohort, such as m308, m399 and m24. Those reads were likely to have been eliminated because the trimmed sequence was too short to map unambiguously to the mm10 reference genome. Consequently, any unique integration sites represented by those sequence reads would have been excluded from the final numbers of unique integration sites identified for the EF1α.γc and PGK.γc cohorts.

An alternative analytical strategy could mitigate this problem by quantifying the number of unique reads for each sample after trimming of the LTR and linker cassette sequences. This remaining sequence, representing the genomic integration, would be a unique mark for the purpose of analysing clonal complexity. Mapping of these genomic sequences is not necessarily essential to the read-out of clonal complexity. Whilst this approach was useful for identifying the dominant clone in m203, it was not applied to all samples because of the results presented in Chapter 4 with respect to NGS analysis of complex barcode libraries, which consisted of over 1 million combinations of random nucleotide sequences. In that chapter, sequencing error prevented the identification of barcode sequences from background within a complex barcode library. If a similar analytical approach to that described in Chapter 4 were applied to the trimmed genomic sequences, it is likely to be similarly impugned by sequencing error, and therefore may only be useful for identifying the presence of dominant clones. Additionally, further biological insights, such as the nature of integration sites
harboured by clones that successfully engrafted in EF1.γc-treated mice compared to PGK.γc-treated mice, could not be explored using such an approach.

For certain samples, particularly m203 and m308, substantially smaller proportions of trimmed sequence reads successfully mapped to the mm10 reference genome. The stringent criteria of tolerating 0 mismatches when reads were aligned using Bowtie2 may have reduced the numbers of trimmed reads that mapped to the genome. The C57Bl/6 strain of the IL2Rγ−/− donor mice may differ from the mm10 reference genome by single nucleotide variants, which may not have been accommodated by the stringent alignment criteria used. An alternative analytical strategy could therefore have used an alternative reference genome, such as the genome of the C57Bl/6NJ strain, the sequence for which is available from the Wellcome Trust Sanger Institute <http://www.sanger.ac.uk/resources/mouse/genomes/>. Structural differences from the mm10 reference genome used, such as chromosomal rearrangements and inversions, could also account for the failure of reads to map. Additionally, fewer unique integration sites may have been detected in certain samples because of the presence of a dominant clone, as was found to be the case for m203. If a dominant clone is rendered unmappable by virtue of short genomic DNA contribution to reads or on account of differences between the donor strain and the reference genome, the associated reads would vastly decrease the proportion of mappable reads.

Conversely, it is also possible that sequencing error may have artificially elevated the number of unique integration sites detected. The questions of whether and to what extent sequencing error may impact upon integration site analysis methods, such as the one used in this chapter, are also unknown. For example, a few errors in a sequence read could cause that read to map erroneously to another locus. It may be
possible to bioinformatically simulate the impact of stochastic sequencing error, for example, by randomly inserting mutations to mapped sequence reads and quantifying how many reads subsequently map to different sites. If sequencing error affects the output of NGS analysis of vector-genomic libraries, however, it would seem likely that each of the 12 libraries analysed in this chapter would be affected in a similar manner. The potential impact of sequencing error on integration site analysis is discussed further in Chapter 6, in light of the findings in Chapter 4 about sequencing error.

Finally, while EF1α.γc-treated mice and PGK.γc-treated mice received equivalent doses of vector-exposed Sca1⁺ BM cells, it is a limitation of this study that the proportion of those cells that were transduced is unknown. This information would be informative for interpreting the integration site analysis of lymphocyte complexity. For example, if a significantly higher proportion of cells were transduced with PGK.γc than with EF1α.γc, then the number of uniquely marked clones that were measured for the PGK.γc cohort may be artificially elevated.

### 3.5.3 Impaired reconstitution of NK cell and B cell subsets following treatment with the EF1α.γc and PGK.γc lentiviral vectors

Compared to mice transplanted with an equivalent cell dose of wild-type Sca1⁺ BM, mice treated with the EF1α.γc and PGK.γc vectors had reduced reconstitution in the B cell and NK cell compartments. For many of the PGK.γc-treated mice these compartments were not present at all. NK cell counts were below the normal range in all patients treated in the French trial of SCID-X1 gene therapy, and in most patients treated in the British trial (Table 1.4; Hacein-Bey-Abina et al. 2010; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011). The presence of T cells, but reduced or absent
NK cells, however, contrasts with what is known about the retention of IL-15 signalling over IL-7 signalling when γc expression is limiting in human cells (Smyth et al. 2007).

The previous analysis of lymphocyte reconstitution achieved with these vectors had detected a low proportion of reconstituted NK cells for both the EF1α.γc and PGK.γc vectors (Ginn et al. 2010). That study, however, had used recipient mice with a different genotype and background strain from those used in the present study. The use of different recipient mice could account for the lack of NK cell reconstitution. Further uncharacterised differences may also exist between the human and murine SCID-X1 phenotypes.

The requirement of γc for B cell development in mice is a point of difference between the human and murine SCID-X1 phenotypes. Although γc expression is not required for B cell ontogeny in humans, it is interesting that the reconstitution of this compartment was reduced in mice that received gene therapy, compared to those treated with wild-type Sca1+ BM cells. In mice treated with the PGK.γc vector, B cell reconstitution was particularly impaired. Interestingly, none of the lymphomas had a B cell phenotype, despite the reduced reconstitution of this compartment.

3.5.3 Limitations of the molecular analyses of lymphomas

There was a discrepancy between the average VCN and the number of vector integration sites detected in 2 of the lymphoma cases. Due to the possibility that those 2 samples may have unidentified integration sites, the possibilities of insertional mutagenesis events at unidentified integration sites cannot be wholly eliminated. The discrepancy is likely to reflect the biases of the conventional LM-PCR method used to identify integration sites. These biases are introduced by the use of the restriction
endonuclease $Tsp509I$ and multiple PCR cycles (Section 1.5.4). If the distance between an integration site and the nearest $Tsp509I$ site is large, then it is unlikely that the site will be detected. Conversely, if the distance between an integration site and the nearest $Tsp509I$ site is small, then it is unlikely to successfully map to the genome. One means of addressing this limitation could be to use restriction endonucleases with different recognition sites. Alternatively, restriction endonuclease-free methodologies have been developed more recently for integration site analysis (Gabriel et al. 2009; Wu et al. 2013).

Furthermore, the molecular analyses of lymphomas do not provide mechanistic insight into the types of somatic genetic lesions caused by replicative stress and how these can cause malignancy. Such causative mechanisms are complex, involving multiple steps that are difficult to isolate and identify, and are beyond the scope of this work.

3.6 CONCLUSIONS FOR CHAPTER 3

The conclusions that can be drawn from this chapter are:

(1) The PGK.γc vector results in reduced overall levels of lymphocyte reconstitution compared to the EF1α.γc vector, with reconstitution skewed towards the T cell compartment;

(2) The PGK.γc vector does not result in a higher rate of lymphomagenesis; and

(3) Analysis of whether the PGK.γc vector results in reduced clonal complexity in the lymphocyte compartment is inconclusive.
CHAPTER FOUR:

GENERATION OF LIBRARIES OF BARCODED VECTORS FOR ANALYSING CLONAL COMPLEXITY AND EVALUATION OF LIBRARY COMPLEXITY USING NEXT GENERATION SEQUENCING

4.1 INTRODUCTION

Retroviral vectors, such as lentiviral vectors and gammaretroviral vectors, have demonstrated great potential as both research tools and gene therapy vectors, particularly for applications of gene therapy targeting the haematopoietic compartment (Sections 1.2.4-5 and 1.3.1). Therapeutic efficacy following retroviral gene delivery to HPCs has been reported in trials of gene therapy for SCID-X1 (Cavazzana-Calvo et al. 2000; Hacein-Bey-Abina et al. 2010; Gaspar et al. 2004; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011), ADA-SCID (Aiuti et al. 2002; Aiuti et al. 2009; Gaspar, Cooray, Gilmour, Parsley, Zhang et al. 2011), CGD (Ott et al. 2006), ALD (Cartier et
al. 2009), WAS (Boztug et al. 2010), leukaemia (Kalos et al. 2011), and attenuation of GVHD (Di Stasi et al. 2011). Analyses of vector integration sites, which function as unique marks of individual gene-marked HPC clones, have the potential to yield important insights into clonal complexity and dynamics following gene therapy. For example, analysis of samples taken at 12–102 months post-transplant from 8 patients treated in the French SCID-X1 trial showed that diversity of reconstituted T cells correlated positively with the dose of genetically modified HPCs received by each patient (Wang et al. 2010). Additionally, the proportion of genetically modified HPCs that contributed to long-term haematopoiesis was estimated as 1%. As knowledge evolves about how many progenitor clones contribute to the differentiation of different haematopoietic lineages and how this changes over time, this may inform definition of optimal cell doses to use during clinical trials.

At present, little is known about how features of vector and protocol design may influence the effective dose of HPCs. For example, the selection of an alternative envelope protein to VSVG can influence transduction efficiency of CD34+ HPCs and hence increase the number of HPCs competent for long-term haematopoiesis (Trobridge et al. 2010). The use of serum-free media and reduced cytokine concentration have been shown to increase transduction efficiency of human CD34+ HPCs, and this may increase the number of progenitor clones contributing to long-term haematopoiesis (Uchida et al. 2011). The research described in Chapter 3 used analysis of unique vector integration sites to analyse whether a less transcriptionally active promoter-enhancer element reduces the dose of Sca1+ BM cells competent for lymphocyte reconstitution in a murine model of SCID-X1, although the results of that analysis were ultimately inconclusive.
Integration site analysis methodologies, such as LM-PCR and LAM-PCR, are presently the gold standard for monitoring clonal diversity. However, these methodologies are associated with biases introduced by the use of restriction endonucleases and many PCR cycles (Section 1.5.4). Such biases limit the potential for quantitative read-outs and can reduce sensitivity for detecting certain clones. For example, it has been shown that amplicon length bias can mask detection of a malignant clone known to be present at 50-fold higher abundance than other clones (Gabriel et al. 2009). Integration site methodologies typically involve multiple steps performed over several days, which may limit how frequently they can be used as a test for monitoring patient samples in the context of a gene therapy trial.

The need for sensitive methods for monitoring of clones arising from individual HPCs is critical in any application of gene therapy targeting the haematopoietic compartment. This has been highlighted following the development of leukaemia in patients treated in gene therapy trials for SCID-X1, CGD and WAS, as well as a non-malignant clonal expansion following gene therapy for β-thalassemia (Sections 1.3.2 and 1.3.6; Hacein-Bey-Abina et al. 2003; Hacein-Bey-Abina et al. 2008; Howe et al. 2008). The development of methods with improved sensitivity and greater quantitative potential for monitoring clonal complexity may facilitate earlier detection of clonal dominance and malignancy, enabling more expedient treatment.

Barcoded vectors, containing a random nucleotide sequence at defined positions, are an attractive alternative approach for analysing clonal complexity. Individual HPCs would be uniquely tagged so long as the barcoded vector stock has sufficiently high complexity. The development of barcoded vectors may offer greater potential for quantifying clonal contributions as well as methodological simplicity, if
minimal PCR cycles are used to amplify the barcode from the genomic DNA of transduced cells for subsequent sequence analysis. Given that doses of transplanted \(\gamma_c^+\)CD34\(^+\) cells in the French and British SCID-X1 trials ranged from 1–22 and 4–8 million cells per kg of body weight, respectively (Hacein-Bey-Abina et al. 2010; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011), it would be necessary for the complexity of a barcoded vector stock to contain millions to tens of millions of combinations of barcode sequences to ensure that each genetically marked HPC clone is uniquely tagged. Analysing the diversity of a highly complex barcode library requires an ability to accurately identify large numbers of unique barcode variants, present at low frequency and in the order of thousands to millions of combinations.

Developments in NGS, particularly the capacity to analyse tens to hundreds of millions of short sequence reads from a single sample, have opened the potential to identify rare barcode variants within a complex mixture of barcode variants that are present at low frequency. Current NGS technologies, however, have higher error rates than traditional Sanger sequencing (Lam et al. 2012; Ratan et al. 2013). This is known to impact on the detection of rare variants, for example, those within a viral population, such that bioinformatic tools are being developed to minimise the detection of false variants (Watson et al. 2013; Wilm et al. 2012). Consequently, sequencing error may result in the false detection of unique barcode variants. Sequencing error may therefore be highly problematic for the analysis of a highly complex library containing thousands to millions of barcodes of unknown sequence identity. For libraries of such complexity, it may be difficult to distinguish between true barcodes and background produced by sequencing error.
One study, which used Illumina NGS technology to analyse a complex barcode library, attempted to address this problem by analysing a portion of the reads that had known sequence identity at a defined region and exploiting the higher abundance of the reads with the correct sequence at that region (Lu et al. 2011). Only approximately 80% of reads contained that known sequence, however, indicating a high proportion of background. Furthermore, the effectiveness of that analytical strategy was not validated empirically. Analyses of complex TCR and antibody repertoires using NGS are an analogous application to barcoding (Briney et al. 2012; Klarenbeek et al. 2010; van Gisbergen et al. 2011). Analysis of known monoclonal TCR sequences, however, has shown that sequencing error produces false TCR sequences, which highlights the need for caution in the interpretation of sequence data for complex TCR repertoires (Nguyen et al. 2011). At present, it remains unknown the degree to which sequencing error impacts upon the analysis of complex barcoded libraries, and whether there is a limit to the degree of complexity that can be analysed using contemporary NGS technology.

This chapter investigates the potential of barcoded lentiviral vectors for use as an alternative to conventional integration site analysis for evaluating clonal complexity. It also explores the feasibility of NGS technologies for analysing the number and relative abundance of different barcode identities within complex libraries of barcoded vectors. Preliminary NGS analyses of a highly complex barcoded plasmid library, in which the barcode was based on a previously reported design (Gerrits et al. 2010), indicated that sequencing error produces a high proportion of background. These analyses were used to inform the design of an optimal barcode, used to generate a second highly complex barcoded plasmid library. Focus shifted towards defining the limit of library complexity that can be analysed using NGS. Monoclonal plasmids that contained known barcode sequences were mixed to produce barcode libraries of
defined complexity comprising known barcode sequences. Whilst the expected barcode sequences in low complexity libraries could readily be distinguished from background, the expected sequences in a moderately complex library could not. This empirical approach to evaluating the impact of sequencing error upon the analysis of complex barcode libraries has enabled insights into the analysable degree of library complexity using NGS and the nature of sequencing error.

4.2 CHAPTER 4 HYPOTHESIS AND AIMS

The hypothesis underlying the investigations in this chapter was:

(1) Barcoded lentiviral vectors offer greater quantitative potential and methodological simplicity for analysing clonal diversity in comparison to conventional integration site analysis.

The aims of this chapter were:

(1) To develop a primer extension method for efficiently generating a complex barcoded plasmid library, using the pEF1α:γc lentiviral construct;
(2) To investigate the feasibility of NGS technologies for analysing the number and relative abundance of barcodes within a complex barcoded plasmid library;
(3) To evaluate the impact of sequencing error upon the analysis of complex barcoded plasmid libraries; and
(4) To define the limit to the degree of library complexity that can be analysed using NGS.
4.3 MATERIALS AND METHODS

All work described in this chapter was performed personally by the author, except where otherwise is indicated.

4.3.1 Generation of complex barcoded plasmid libraries

4.3.1.1 Primer extension methodology to construct double-stranded barcode inserts

A primer extension method was developed to construct double-stranded barcode inserts for cloning into the single NsiI site of pEF1α.γc, described in Section 3.3.1.1 (Figure 4.1). Oligonucleotides containing random nucleotides at defined positions were synthesised (Table 4.1). For some barcode designs, the known sequences comprised whole or part of the adaptor sequences for the Illumina or Sequencing by Oligonucleotide Ligation and Detection (SOLiD; Life Technologies) platforms. Barcode oligonucleotides, Barcode design 1, Illumina-specific barcode and SOLID-specific barcode, were annealed with their respective primers for second strand synthesis by heating the oligonucleotides, in the presence of annealing buffer (Table 2.4) and in a final volume of 16 μL, to 95°C for 5 minutes in a heating-block. The heating-block was then switched off to cool the samples slowly to room temperature. This step created an NsiI-compatible end at the 3′ end of the barcode oligonucleotides.

The complementary strands of barcode inserts were synthesised using the 3′→5′ exo− Klenow Fragment (New England Biolabs). A final volume of 20 μL contained 16 μL of the annealed oligonucleotides, reaction buffer and 10 U Klenow Fragment. Reactions were incubated at 37°C for 15 minutes. An NsiI-compatible end was generated at the 5′ end of the barcode oligonucleotide by cleavage with PstI (New
Figure 4.1: Schematic diagram of primer extension method used to construct a barcode insert. The barcode oligonucleotide Barcode design 1 contained 12 degenerate nucleotides at defined positions as indicated and comprised the template strand for second strand synthesis. A primer with homology to the 3'-end of the barcode oligonucleotide was annealed, generating a $\text{NsiI}$-compatible overhang, and extended to produce the second strand of the insert. This product was cleaved using $\text{PstI}$, which recognises the double-stranded “CTGCAG” site highlighted in green, producing a $\text{NsiI}$-compatible overhang at the 5'-end of the barcode oligonucleotide. The $\text{PstI}$-digested barcode insert was then cloned into the single $\text{NsiI}$ site of the pEF1α.γc lentiviral plasmid construct. EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; nt, nucleotide.
England Biolabs). A final volume of 200 μL contained reaction buffer and 2,000 U
\textit{PstI}. Reactions were incubated at 37°C for 15 minutes.

Barcode DNA was purified by column-based purification (Section 2.3.3.4). Functionality of \textit{NsiI}-compatible sticky ends was confirmed by self-ligation of 1 μL of the barcode insert using 0.5–0.75 U of T4 DNA Ligase (Promega; Section 2.3.5.2). The ligation products were visualised following separation by polyacrylamide gel electrophoresis (Section 2.3.3.3).

\textbf{Table 4.1: Oligonucleotide sequences used for construction of barcode inserts}

<table>
<thead>
<tr>
<th>Oligonucleotide*</th>
<th>Sequenceb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcode design 1</td>
<td>5'-ATAGGTTCCATGACATCGTATCGACTGAGGNAACGNVNGATNNACANVNGTNTACNNNGTTGTAACACGAAGGCGCTGC-3'</td>
</tr>
<tr>
<td>Barcode design 1 second strand</td>
<td>5'-GGCCCTTCGTGTTACAA-3'</td>
</tr>
<tr>
<td>Illumina-specific barcode</td>
<td>5'-ATAGTCCATGACATCGTACCTGAGCAGCCTCTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNGGTGCACGGGTGCCTGCA-3'</td>
</tr>
<tr>
<td>Illumina-barcode second strand</td>
<td>5'-GGCACCCCGTGCAC-3'</td>
</tr>
<tr>
<td>SOLiD-specific barcode</td>
<td>5'-ATAGTCCATGACATCGTACCTGAGCAGCCTCTTCTATGGGCAGTCGGTGCTNNNNNNNNNNNNNNNNNACGCCTTGCCGTACAGCAGCTGCA-3'</td>
</tr>
<tr>
<td>SOLiD-barcode second strand</td>
<td>5'-GCTGCTGTACGCAAGGCG-3'</td>
</tr>
</tbody>
</table>

*All oligonucleotides were purchased from Sigma-Aldrich.

bLocations of degenerate nucleotides within the barcode oligonucleotides are indicated using bold and italics.
4.3.1.2 Cloning of double-stranded barcode inserts into pEF1α.γc for generating a complex barcoded plasmid library

In order to ensure maximum power in a gene therapy context, it is necessary for each of the transduced HPCs in a preclinical or clinical application to be labelled with a unique barcode. Assuming a 10 kg patient in a clinical trial were to receive $10 \times 10^6$ cells per kg of body weight, of which 1% give rise to lymphocyte reconstitution, it would be necessary for the complexity of a barcoded plasmid library to exceed 1 million combinations (Section 4.1). The generation of such a highly complex barcoded plasmid library requires a high efficiency at all cloning stages. The following method was developed to maximise the complexity of barcoded pEF1α.γc libraries. The lentiviral construct pEF1α.γc was linearised using a 20-fold unitary excess of NsiI (New England Biolabs) to plasmid DNA. Following heat-inactivation of NsiI, the linearised vector backbone was dephosphorylated using Antarctic Phosphatase (New England Biolabs), which was also heat-inactivated before column-based purification of the vector DNA (Section 2.3.3.4). The NsiI-linearised pEF1α.γc vector was ligated with a 10-fold molar excess of the barcode insert (Section 2.3.5.2), after which the NsiI site was not reconstituted. This enabled digestion of the ligation product with NsiI to linearise vector molecules that re-ligated without the barcode insert. Linearised molecules lacking the barcode insert were not taken up by bacteria during transformation, and thus vector molecules containing the barcode insert were selected. This method ensured the plasmid library did not contain any vector molecules that lacked the barcode insert. The ligation product was ethanol precipitated to remove salts before transformation (Section 2.3.3.6).
Electrocompetent SURE cells (Table 2.2; transformation efficiency of $\geq 1 \times 10^{10}$ transformants per $\mu$g of supercoiled pUC18 DNA) were selected for use in the transformation of the ethanol precipitated-ligation product, in order to maximise the number of transformed bacteria and hence the complexity of the barcoded plasmid library purified from those bacteria after they were cultured. Large scale transformations of electrocompetent SURE cells were performed to prepare stocks of barcoded pEF1α.γc plasmid DNA (Section 2.3.4.2). Approximately 250 $\mu$L of bacteria were transformed with 830 ng of ligated DNA. A small portion of transformed cells were diluted and spread on LB plates for the purpose of estimating transformation efficiency. The remaining cells were used as a starter culture for an overnight culture for a large-scale (Gigaprep; Qiagen) isolation of plasmid DNA.

**4.3.1.3 Small-scale validation of library complexity**

For all barcoded plasmid libraries, a representative sample of single colonies were picked from the bacterial plates used to estimate transformation efficiency (20 colonies for the library containing Barcode design 1 and over 100 colonies for each of the libraries containing the Illumina-specific and SOLiD-specific barcodes). Plasmid DNA was purified (Section 2.3.2.1) and Sanger-sequenced (Section 2.3.7) for preliminary validation of library complexity and sequencing of a representative proportion of unique barcodes.

**4.3.2 Next generation sequencing of barcode libraries**

The barcode regions in libraries of barcoded pEF1α.γc were PCR-amplified for sequencing on either the Illumina or SOLiD platforms using primers that introduced the necessary adaptor sequences required for sequencing (Table 4.2). These primers also
Table 4.2: Oligonucleotide sequences used for amplification of barcodes as preparation for next generation sequencing

<table>
<thead>
<tr>
<th>Barcode design</th>
<th>Oligonucleotide&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sequence&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Barcode design 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD_Lenti_Fwd1</td>
<td>5'-AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTTCCGATCTAATAGCTACCCGACGCTTGACGAT-3'</td>
<td></td>
</tr>
<tr>
<td>CD_Lenti_Fwd8</td>
<td>5'-AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTTCCGATCTGCTCTACCCGACGCTTGACGAT-3'</td>
<td></td>
</tr>
<tr>
<td>CD_Lenti_Fwd15</td>
<td>5'-AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTTCCGATCTCGTAGCTACCCGACGCTTGACGAT-3'</td>
<td></td>
</tr>
<tr>
<td>CD_Lenti_Prev</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td><strong>Illumina-specific barcode</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal forward primer</td>
<td>5'-AATGATACGGCGACCACCAGATCTACACTCTTTCCCTACACGACGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #1</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #2</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #3</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #4</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #5</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #6</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #7</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td>Illumina Reverse #8</td>
<td>5'-CAAGCAGAAGACGGCATACGAGCTCTTTCCGATCTGTAAGTCATTGGTCTTAAAG-3'</td>
<td></td>
</tr>
<tr>
<td><strong>SOLiD-specific barcode</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOLiD forward primer</td>
<td>5'-CCACTACGGCTCCGCTCTCTCTATGGGACGCTTG-3'</td>
<td></td>
</tr>
<tr>
<td>SOLiD Reverse #7</td>
<td>5'-CTGCCCCCGGGGCTCTATTCTCTTGATGCTTGCTGTACGGGCAACCCGTCACC-3'</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>All oligonucleotides were purchased from Sigma-Aldrich.

<sup>b</sup>Sequences of sample indexes, used for multiplexing of different samples, are indicated using bold and italics.
introduced a sample index for identifying samples, enabling different samples to be multiplexed in a single sequencing run.

For barcode design 1, amplification used Pfu polymerase (2–3 U/µL; Promega) and 62 cycles of PCR. Each 50 µL reaction contained 100 ng of template plasmid DNA, reaction buffer, the 4 dNTPs (0.2 mM), forward and reverse primers (0.8 mM), and 0.6 µL Pfu polymerase. Cycling conditions consisted of an initial denaturing step at 95°C for 5 minutes; followed by 2 cycles of template denaturing (95°C for 30 seconds), primer annealing (60°C for 30 seconds) and extension (72°C for 30 seconds); then 60 cycles of template denaturing (95°C for 30 seconds), primer annealing (80°C for 30 seconds) and extension (72°C for 30 seconds); followed by a final extension step (72°C for 5 minutes). Samples were purified using column-based purification (Section 2.3.3.4). For the subsequent barcode designs, the PCR amplification was reconfigured to minimise the possibility of PCR error, through use of the polymerase with the highest available fidelity, Phusion (2 U/µL; Thermo Scientific), and minimal amplification cycles. Each 50 µL reaction contained 200 ng of template plasmid DNA, reaction buffer, the 4 dNTPs (0.2 mM), forward and reverse primers (0.5 mM), and 0.25 µL Phusion polymerase. Cycling conditions consisted of an initial denaturing step at 95°C for 5 minutes; followed by 10 cycles of template denaturing (95°C for 30 seconds), primer annealing (55°C for 30 seconds) and extension (72°C for 30 seconds); followed by a final extension step (72°C for 5 minutes). Samples were purified using column-based purification (Section 2.3.3.4).

Sequencing on the Illumina platform was conducted on a HiSeq2000 instrument using either 50 or 100 base single-end reads. Sequencing of barcode design 1 used 100 base single-end reads (Australian Genome Research Facility). Sequencing of the
Illumina-specific barcode used 50 base single-end reads (Beijing Genomics Institute). Sequencing of the SOLiD-specific barcode was conducted on a 5500xl instrument (Victor Chang Cardiac Research Institute).

4.3.3 Data analysis

Bioinformatic analyses used a combination of standard UNIX tools for string manipulation, the MySQL relational database for sequence counting and customised Perl scripts for data manipulation, filtering, clustering and error analysis.

All customised Perl scripts used in this chapter were written by Mr Jeffrey Deakin, except where indicated as written by the author, and are reproduced in Appendix 1.

4.3.3.1 Data analysis for the initial barcode design

Raw sequence reads were processed using the program barcode-parse-initial.pl, which reduced the 100-base reads to the positions of the reads that corresponded to the barcode and the quality values for each of those barcode positions. The program barcode-filter-initial.pl was used to eliminate sequence reads where the quality score of any of the barcode positions was below a given threshold. A quality Phred score of 30, corresponding to a 99.9% probability of accurate base-calling, was typically used. Unique barcode sequences were counted and then listed in order of decreasing abundance using a simple MySQL query.

The clustering program, cluster-initial.pl, was designed to assume no prior knowledge of genuine versus erroneous barcodes. Barcodes were processed in a
hierarchical fashion using the correct order produced by MySQL. For each pair of barcodes under comparison the Hamming distance, which refers to the number of different positions between two different sequences, was calculated (Hamming 1950). Briefly, all detected sequences were compared with the first-most abundant barcode and their counts were added to those of the first-most abundant barcode if they differed by 1 position, which was determined by calculating the Hamming distance of each sequence relative to the first-most abundant barcode. The process was repeated for the second-most abundant barcode and then the third-most abundant barcode and so on, until all remaining barcodes had been processed.

4.3.3.2 Data analysis for the Illumina-specific barcode design

For the Illumina-specific barcode, raw sequence reads obtained from the Illumina HiSeq2000 were initially filtered for the known sequence, “GGTGCACGGGTGCC” at positions 17-30. This facilitated elimination of errors caused by nucleotide insertion or deletion. Filtered reads were trimmed to positions 2-16 of the barcode using barcode-parse.pl. Reads where any of the barcode positions had a Phred score below 30 were filtered using barcode-filter.pl. Unique barcode sequences were counted using MySQL and clustering was performed using cluster.pl.

4.3.3.3 Data analysis for the SOLiD-specific barcode design

For the SOLiD-specific barcode, raw sequence reads obtained from the SOLiD 5500xl were filtered based on the presence of the first 10 nt of the internal adaptor sequence, “ACGCCTTGCC”, at positions 16-25. This facilitated elimination of errors caused by nucleotide insertion or deletion. Subsequent analyses involved counting unique barcode sequences using MySQL and clustering using cluster.pl (Section 4.3.3.2). Phred score
quality filtering was not performed due to the different format of the sequencing data output from the SOLiD 5500xl, and the development of scripts for such analysis was not justified following the analysis of this preliminary dataset.

4.3.3.4 Analyses of empirical error frequencies and error types

For the Illumina-specific barcode, empirical error frequencies were assessed using mismatch-barcode.pl, written by the author, and the sequence dataset for a sample containing 1 known barcode. This script compared each sequence read of that sample with the known barcode sequence in a position-wise manner for the barcode region, and counted the total number of errors at each position. This analysis was performed using both the filtered and unfiltered sequence reads. For the initial barcode design, analysis of empirical error frequencies was performed by the Bioinformatics Service of the Australian Genome Research Facility.

Analysis of 1-mismatch errors was performed using error-analysis.pl, which compared the 1-mismatch error sequences generated by cluster.pl to the expected barcode sequences from which they differed by 1 mismatch. This analysis was applied to defined libraries containing known sequences and assumed that the sequences that differed from expected barcodes by 1 mismatch were generated by errors at the mismatch position. Whilst this analysis provided insight into the type and location of single nucleotide substitution-like errors, it was uninformative about the systematic errors that produced putatively false barcodes containing 6 or more errors.
4.3.3.5 Analyses of sequence characteristics

GC content for each of the barcode sequences in the defined barcode libraries was calculated using gatc-string.pl, which was written by the author. Minimum Gibbs Free Energy values for the 100 known barcode sequences and the whole barcode amplicons were calculated using UNAfold (Markham & Zuker 2008). Barcodes not included in the 100-barcode library, but which were detected within the top 120 unique barcodes for the Illumina-specific 100-barcode library, were compared to each of the 100 expected barcodes using Hamming distances calculated by 100-noise-hamming-distance.pl. Multiple sequence alignments and analyses of conserved regions were conducted using BioEdit (Version 7).

4.3.4 Empirical assessment of individual barcode representation within the Illumina-specific 100-barcode library

The presence and relative abundance of selected individual barcode sequences within the mixture of monoclonal plasmids containing 100 Illumina-specific barcode sequences were estimated using qPCR assays devised for 10 specific barcode sequences. Selection of individual barcodes for qPCR analysis was based on the relative frequencies of expected barcodes within the Illumina-specific 100-barcode library, as detected by the first sequencing run on the HiSeq2000. The individual barcodes selected comprised the 3 most abundant expected barcodes, 3 barcodes with mid-levels of abundance and the 3 undetected barcodes, based on the abundances of barcodes from the first sequencing run. In addition, 2 barcodes not included in the Illumina-specific 100-barcode library, yet unexpectedly detected at the highest abundance by the HiSeq2000, were selected for qPCR analysis.
Each qPCR assay used a primer designed to specifically bind to the barcode sequence and a vector-binding primer designed to produce an approximately 120-bp product (Table 4.3). The 3' end of the barcode-specific primers contained 11 nt that specifically bound to the barcode, such that the remaining 5 nt of the barcode sequence served as template for amplification. These 5 nt functioned as a 'signature' for each barcode sequence and were used to confirm the specificity of amplification when PCR products were Sanger-sequenced. Primers Top1_recessed, Top2_recessed and Top3_recessed amplified the first-most, second-most and third-most abundant barcodes, respectively. Primers Middle1_recessed, Middle2_recessed and Middle3_recessed amplified the forty-ninth-most, fiftieth-most and fifty-first-most abundant barcodes, respectively. Primers Bottom1_recessed, Bottom2_recessed and Bottom3_recessed amplified 3 barcodes that were undetected in the Illumina-specific 100-barcode library. Since the monoclonal plasmid which contained the barcode sequence amplified by Bottom2_recessed actually contained 2 barcode inserts, an additional primer, Primer Bottom2_control_recessed, was designed to amplify that second barcode. For the sake of consistency, all of these primers were used in combination with the Vector-specific primer.

For each qPCR assay, a $1 \times 10^{10}$-copy standard of the monoclonal plasmid containing the barcode of interest was prepared and then serially diluted 1 in 10, such that $1 \times 10^3$-copy to $1 \times 10^9$-copy standards were produced for constructing a standard curve. Plasmids containing the 2 putatively false barcodes that were selected for analysis were constructed for the purpose of generating a qPCR standard curve. Ten million-copy standards were prepared for each of the samples that were analysed, which included the 100-barcode library, a 10-barcode mixture containing the barcode of interest, the monoclonal plasmid containing the barcode of interest, and 3
Table 4.3: Oligonucleotide sequences used for validation of the Illumina-specific 100-barcode library

<table>
<thead>
<tr>
<th>Oligonucleotide(^a)</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector-specific</td>
<td>5'-CCCATGTTCACCCCTAAAGCCTGAA-3'</td>
</tr>
<tr>
<td>Top1_recessed</td>
<td>5'-CTACACGACGCTCTTCCGATCTTATTATTACAC-3'</td>
</tr>
<tr>
<td>Top2_recessed</td>
<td>5'-ACCCGTGCACCGCAGAGAGGC-3'</td>
</tr>
<tr>
<td>Top3_recessed</td>
<td>5'-CACGACGCTCTTCCGATCTAATTCTTTTGGC-3'</td>
</tr>
<tr>
<td>Middle1_recessed</td>
<td>5'-CAGGCACCCTGCACCTTTCTCCTTCTTCTC-3'</td>
</tr>
<tr>
<td>Middle2_recessed</td>
<td>5'-CACCGTGCACCACGAGAGGC-3'</td>
</tr>
<tr>
<td>Middle3_recessed</td>
<td>5'-CCCACCCCTGCACCTTTCTCTCCC-3'</td>
</tr>
<tr>
<td>Bottom1_recessed</td>
<td>5'-GACGCTCTTCCGATCTTGTCTGAC-3'</td>
</tr>
<tr>
<td>Bottom2_recessed</td>
<td>5'-GGACCACCCCTGCACCACTCTGCTGAC-3'</td>
</tr>
<tr>
<td>Bottom2_control_recessed</td>
<td>5'-GCACCCCTGCACCCAGACAGACACAC-3'</td>
</tr>
<tr>
<td>Bottom3_recessed</td>
<td>5'-GGACCCGCACCGTGAGCCTTGGTGAT-3'</td>
</tr>
<tr>
<td>Spike_recessed</td>
<td>5'-GACGCTCTTCCGATCTATGGCTAGG-3'</td>
</tr>
<tr>
<td>Vector-specific_ReverseOri</td>
<td>5'-GGCTAAGATCTACAGCTGTTC-3'</td>
</tr>
<tr>
<td>False1_recessed</td>
<td>5'-GGACCCCGTGCACCCAGCTGATTC-3'</td>
</tr>
<tr>
<td>False2_recessed</td>
<td>5'-GGACCCCGTGCACCCATTTGAGTCA-3'</td>
</tr>
<tr>
<td>False3_recessed</td>
<td>5'-GGACCCCGTGCACCTACCTACTT-3'</td>
</tr>
<tr>
<td>False4_recessed</td>
<td>5'-GCACCCCGTGCACCCAGGTAAG-3'</td>
</tr>
<tr>
<td>False5_recessed</td>
<td>5'-CCGTGCACCCCTGACAGCAGC-3'</td>
</tr>
<tr>
<td>False6_recessed</td>
<td>5'-GGACCCCGTGCACCCAAATGTACAG-3'</td>
</tr>
</tbody>
</table>

\(^a\)All oligonucleotides were purchased from Sigma-Aldrich.

monoclonal plasmids containing different barcodes. The abundance of the barcode of interest in the 100-barcode library sample was calculated relative to the sample comprising the monoclonal plasmid that contained the barcode of interest.
Reactions used the SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich) and contained barcode-specific and vector-binding primers (0.2μM). Reactions were performed in a RotorGene 6000 (Corbett Life Science), and cycles consisted of an initial denaturing step (95°C for 10 minutes), followed by 25 cycles of template denaturing (95°C for 20 seconds) and primer annealing and extension (72°C for 20 seconds), and a final extension step (72°C for 7 minutes). Each assay had a reaction efficiency of greater than 86% and R² of ≥ 0.99. The specificity of barcode amplification was confirmed by Sanger-sequencing of the PCR product that was produced when the 100-barcode library served as template for amplification, for the samples that yielded an amplification product.

In order to estimate the limit of sensitivity of this approach for detecting specific barcodes, a barcode sequence that was absent from the 100-barcode library was spiked into the 100-barcode library in known proportions: 0.1%, 0.5%, 0.75%, 1%, 1.25%, 2.5%, 5%, 10% and 20%. The spiked-in barcode was quantified by qPCR using the recessed primer Spike_recessed and the Vector-specific primer to produce a standard curve (Table 4.3). The specificity of amplification of the spiked-in barcode sequence was validated by Sanger-sequencing of the products yielded when the mixtures containing 0.1% and 1% of the spiked-in barcode were used as template for amplification.

Additionally, non-quantitative PCR was performed to detect the presence of 4 additional individual putatively false barcodes, which were within the 100 most abundant barcodes detected in the 100-barcode library following both Illumina sequencing runs. A similar recessed primer strategy was employed to design primers specific for these putatively false barcodes (Table 4.3). These primers were used with
either the Vector-specific primer or Vector-specific_ReReverseOri primer, to accommodate the possibilities of either forward or reverse barcode orientations of the putatively false barcodes, if present.

For non-quantitative PCR, a 25 μL reaction contained 1.25 U Taq DNA polymerase (Roche), reaction buffer, the 4 dNTPs (0.2 mM) and a false barcode-specific and a vector-binding primer (0.2 μM; Table 4.3). Reactions were performed in an Eppendorf Mastercycler Gradient PCR machine and cycles consisted of a denaturing step (95°C for 10 minutes), followed by 25 cycles of template denaturing (95°C for 20 seconds), primer annealing and extension (72°C for 20 seconds), followed by a final extension step (72°C for 7 minutes).

4.3.5 Statistical analysis

Comparisons between observed and expected proportions of errors at barcode positions and each of the possible types of substitution errors were made using χ-squared tests (GraphPad Prism Version 5).

4.4 RESULTS

4.4.1 Construction and validation of an initial barcoded pEF1α.γc lentiviral plasmid library

An 82 bp barcode insert containing 12 degenerate positions, barcode design 1 (Figure 4.1), was based on a previous design for a vector barcode (Gerrits et al. 2010). Barcode design 1 was constructed using a primer extension methodology for bidirectional insertion into the single NsiI site of the lentiviral plasmid construct pEF1α.γc. The efficiency of barcode insert construction was validated at each stage of the method.
With 4 possible nucleotides at each of the 12 degenerate positions and 2 possible orientations for the insert, the theoretical maximum complexity of the barcoded pEF1α.γc library generated using barcode design 1 exceeded 33.5 million possible combinations of barcode tags. In practice, however, library complexity was restricted to the number of bacteria transformed with barcoded pEF1α.γc molecules (Section 4.3.1.2). When this initial barcoded pEF1α.γc library was generated, the transformation efficiency was estimated as $5.3 \times 10^6$ transformants per μg ligation product. One microlitre of pooled bacterial culture contained approximately 400 transformed bacteria after 1 hour of culturing after transformation, in a total of 11 mL of pooled culture. Therefore, the actual complexity of the barcode library was approximately 4.4 million combinations of barcodes.

Sanger sequencing of 20 individual clones enabled small-scale preliminary validation of library complexity and confirmed the unique barcode identity for those 20 clones (Figure 4.2B).

### 4.4.2 Analysis of the initial barcoded pEF1α.γc lentiviral plasmid library using Illumina sequencing

#### 4.4.2.1 Preparation of barcode amplicons for next generation sequencing

The barcode region in samples representing the complex barcode library and a control sample containing 2 known barcodes was amplified using primers that introduced the Illumina adaptor sequences (Section 4.3.2). The samples selected for analysis consisted of the complex barcoded pEF1α.γc plasmid library, genomic DNA from HEK-293 cells transduced with pEF1α.γc vector produced from the complex library, and a 2-barcode sample consisting of 2 monoclonal barcoded plasmids.
Figure 4.2: Validation of barcode insert construction and preliminary validation of library complexity. (A) Step-wise validation of the primer extension methodology for constructing the barcode insert. A sample of the barcode insert was electrophoresed in 10% polyacrylamide following each stage of barcode insert construction to confirm the efficiency of each stage. (B) Proportions of nucleotide representation in the barcode positions of 20 representative monoclonal barcodes. After ligation of the barcode insert into the EF1α.γc lentiviral construct and transformation into bacteria, the barcode region in 20 individual colonies was Sanger sequenced to provide preliminary evidence of barcoded EF1α.γc library complexity. The underlined positions in the pictogram represent the barcode positions from those 20 colonies and the size of each letter depicts the relative proportions of each nucleotide. The pictogram was created using the WebLogo tool at <http://weblogo.berkeley.edu/logo.cgi>. EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; oligo, oligonucleotide; ss, single-stranded.
A total of 5,274,931 sequence reads were obtained for the 2-barcode sample, 8,576,452 reads for the complex plasmid library and 4,343,847 reads for the genomic sample representing vector stock.

4.4.2.2 Analysis of a sample containing 2 known barcodes

Analysis was initially performed using the 2-barcode sample so that the known barcodes could function as reference sequences. Although this sample contained 2 barcodes, 27,284 barcodes were counted (Table 4.4). The distribution of the 500 most abundant barcodes indicated that the 2 expected barcodes occurred at over 60-fold higher frequency than background (Figure 4.3A). Although there were only 2 barcodes detected at high frequency, a large number of barcodes occurred at low frequency, and it was likely that this background was due to sequencing error (Figure 4.3B). Whilst over 90% of all sequence reads could be accounted for by the 2 expected barcodes (Table 4.4), it was clear that there was a high proportion of background that would be problematic for analysing libraries with higher complexity. Therefore, the 2-barcode sample was used to develop bioinformatic strategies for reducing background.
Table 4.4: Effect of bioinformatic strategies on the reduction of background detected in a sample containing 2 known barcodes

<table>
<thead>
<tr>
<th>Bioinformatic strategy</th>
<th>Total sequence reads</th>
<th>Total unique barcodes</th>
<th>Proportion of barcode “CTGGCCGTTTT”</th>
<th>Proportion of barcode “TGGGGCGTTTCAT”</th>
<th>Proportion of background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting unique barcodes</td>
<td>5,274,931</td>
<td>27,284</td>
<td>50.1%</td>
<td>40.4%</td>
<td>9.5%</td>
</tr>
<tr>
<td>Quality filtering (Q30)</td>
<td>930,252</td>
<td>162</td>
<td>89.3%</td>
<td>8.6%</td>
<td>2.1%</td>
</tr>
<tr>
<td>Clustering, with 1 mismatch</td>
<td>5,274,931</td>
<td>13,311</td>
<td>52.4%</td>
<td>43.1%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Q30 filtering and clustering with 1 mismatch</td>
<td>930,252</td>
<td>63</td>
<td>89.5%</td>
<td>8.6%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>
Chapter 4: Generation and analysis of barcoded vectors

Figure 4.3: Analysis of a sample containing 2 defined barcode sequences.  
(A) Relative abundance of the 500 most abundant barcode sequences detected in the 2-barcode sample.  
(B) Distribution of the number of unique barcodes detected at each level of abundance, from an abundance of 1 to an abundance to over 1 million, for barcode sequences detected in the 2-barcode sample.
4.4.2.3 Development of bioinformatic strategies for reducing background

The potential of 3 bioinformatic strategies for reducing background was explored:

(1) Quality filtering;

(2) Homology-based clustering with 1 mismatch; and

(3) The combination of quality filtering and clustering with 1 mismatch.

The quality filtering approach involved filtering out reads where the Phred score, or quality value, for any of the barcode positions was below 30 (Q30), which corresponds to a 99.9% probability that the base call was correctly made during sequencing. The clustering approach was based on an observation that the 10 putatively false barcodes that were most frequently detected in the 2-barcode sample contained single mismatches. The assumption underlying the clustering strategy was that any barcodes differing by 1 position from the relatively frequent barcodes were more likely to represent sequencing error than genuine barcodes. This assumption was considered legitimate because the encoded library complexity, at over 33.5 million, outnumbered the anticipated plasmid library complexity, at 4.5 million, by 7.4-fold. The degree of encoded degeneracy was therefore sufficient to tolerate 1 mismatch. Assumptions about the expected sequence identities of any of the barcodes in a sample do not underlie the clustering approach. Therefore, this analysis could be refined using the 2-barcode sample for subsequent application to the complex library of unknown sequence identity.

Background was successfully reduced to below 2% for the 2-barcode sample using a combination of quality filtering and clustering with 1 mismatch, and the total number of unique barcodes detected was reduced to 63 (Table 4.4). However, the
distribution of the relative abundance of each detected barcode did not yield a clear
distinction between the expected barcodes and background, indicating that it would not
be possible to distinguish between genuine barcodes and background on the basis of
frequency (Figure 4.4). Interestingly, there was a discrepancy in the relative abundance
of the 2 expected barcodes after quality filtering and clustering, with barcode
“CTGGCCGGTTTT” detected at 10.4-fold higher abundance than barcode
“TGGGGCGTTCAT”. This may imply that the reads of barcode “TGGGGCGTTCAT”
were of lower quality. Whilst quality filtering and background successfully reduced
overall levels of background, the remaining background ultimately could not be
distinguished readily from the 2 expected sequences in the 2-barcode sample.

4.4.2.4 Analysis of the complex libraries

Even though Q30 quality filtering and clustering with 1 mismatch did not successfully
resolve complexity of the 2-barcode sample, these analytical strategies were
nevertheless applied to the samples representing the complex barcoded pEF1α.γc
plasmid library and barcoded EF1α.γc vector stock. Over 30,000 unique barcode
sequences were detected in both samples after Q30 quality filtering and clustering with
1 mismatch (Table 4.5). However, the distribution of the relative abundance of
barcodes displayed a gradual tapering-down pattern, suggesting a clear distinction
between true barcodes and background would be absent (Figure 4.5). Notably, Q30
filtering eliminated 99.4% of reads and 94.3% of reads of the plasmid and genomic
samples, respectively. This indicates a large portion of sequence reads for these
samples had insufficient quality for this application.
Figure 4.4: Abundance of unique barcodes detected by Illumina next generation sequencing in a sample containing only 2 known barcodes, after Q30 quality filtering and clustering with 1 mismatch. (A) Relative abundance of the 500 most abundant barcode sequences detected in the 2-barcode sample. (B) Distribution of the number of unique barcodes detected at each level of abundance, from an abundance of 1 to an abundance to over 1 million, for barcode sequences detected in the 2-barcode sample.
Figure 4.5: Abundance of unique barcodes detected by Illumina next generation sequencing analysis of highly complex samples. (A) Relative abundance of the 500 most abundant barcode sequences detected in the sample representing the complex barcoded plasmid library, after Q30 quality filtering and clustering with 1 mismatch. (B) Relative abundance of the 500 most abundant barcode sequences detected in the genomic sample representing vector stock produced from the complex barcoded plasmid library, after Q30 quality filtering and clustering with 1 mismatch.
### Table 4.5: Analysis of unique barcodes detected in samples representing the complex plasmid library and vector stock

<table>
<thead>
<tr>
<th></th>
<th>Complex plasmid library</th>
<th>Complex vector stock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequence reads</td>
<td>8,576,452</td>
<td>4,343,847</td>
</tr>
<tr>
<td>Total sequence reads after Q30 filtering</td>
<td>55,598</td>
<td>245,968</td>
</tr>
<tr>
<td>Discarded reads as a proportion of total reads</td>
<td>99.4%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Total unique barcodes after Q30 filtering and clustering</td>
<td>41,584</td>
<td>33,215</td>
</tr>
<tr>
<td>Number of unique barcodes detected fewer than 5 times</td>
<td>41,405</td>
<td>17,831</td>
</tr>
<tr>
<td>Number of unique barcodes detected once</td>
<td>35,802</td>
<td>8,999</td>
</tr>
</tbody>
</table>

There were 2 barcodes detected at almost 1 log-fold higher abundance than other barcodes in the plasmid sample, and 2 barcodes detected at almost 2 log-fold higher abundance than other barcodes in the genomic sample. For the 500 most abundant barcodes detected in both samples, the relative frequencies of barcodes gradually tapered down after those 2 high frequency barcodes, and there were a large number of barcodes detected at low frequency (Figure 4.5). Of the unique sequences detected in both samples, 99.6% and 53.7% were detected fewer than 5 times and 86.1% and 27.1% were detected once in the plasmid and vector samples, respectively (Table 4.5). Taken together, these results indicate it would not be possible to distinguish between true barcodes and background produced by sequencing error on the basis of frequency.

### 4.4.2.5 Analyses of empirical error frequencies and quality of sequence reads

In order to better understand the nature of the sequencing error that was causing background, further analyses were conducted to characterise the error frequencies and
qualities of sequence reads for this sequencing run. Given that the sequence identity for each position of the sequence reads was known, with the exception of the 12 barcode positions, it was possible to calculate the frequency of errors at those known positions across the length of the sequence read for all sequence reads. The distribution of error frequencies revealed that error frequencies were relatively low, at below 0.1%, for the first 35 nt, with the exception of the first nt (Figure 4.6A). The error frequency at the first nt was high, at 23.3%. After the first 35 nt, the error frequencies steadily increased and were in the order of 5% in the region where the barcode was located. The distribution of average quality values across the length of the sequence read also indicated that the probability of accurate base-calling was higher towards the front end of the sequence reads, with average Phred scores above 30 for the first approximately 35 nt (Figure 4.6B). For each position in the sequence read, the ratio of the observed error frequency to the probability of incorrect base-calling, calculated using the average quality values for each position, was around 2 log-fold above 1 (Figure 4.6C). This implies that predictions about the accuracy of reads for this dataset overestimated the actual accuracy.

A 33 nt homogeneous sequence at the front-end of sequence reads was likely to have impaired the platform’s methods for registering individual bridge amplification clusters (personal communication from Dr Gabriel Kolle, Illumina). The Illumina platform’s algorithms for identifying individual clusters that have arisen from individual molecules relies on sequence heterogeneity at the front-end of reads. This sequence homogeneity may have contributed to the unexpectedly high error frequency. Increased heterogeneity at the front-end of the sequence read, therefore, may have facilitated improved cluster registration and base-calling. It was also possible that the relatively high cluster density for this sequencing run, at 773,000 ± 69,000 clusters per
Figure 4.6: Analyses of the distribution of error frequencies and read qualities for each position in sequence reads. (A) The average frequency of errors at each position, calculated using the sequence reads from all samples and excluding any barcode positions where the identity of a position was unknown. The location of the barcode region, in forward orientation (positions 40-67) or reverse orientation (positions 55-81), is indicated by the blue lines. (B) Average Phred quality for each position in the sequence read. (C) Ratio of the observed frequency of errors to the predicted probability of an incorrect base-call at each position. The predicted probability of an incorrect base-call was calculated using the average quality value and the conversion Quality = -log_{10}(error).
mm\(^2\), reduced the quality of base-calling. Drawing on these insights into the Illumina sequencing technology, the barcode was redesigned in a platform-specific manner in order to take advantage of the highest quality Illumina data. Since a similar configuration would be favourable for sequencing using the SOLiD platform (personal communication from Dr Bennett Shum, Life Technologies), the barcode was also redesigned in a manner specific for that technology.

### 4.4.3 Design, construction and preliminary validation of platform-specific barcoded pEF1\(\alpha,\gamma\)c lentiviral plasmid libraries

Redesigned barcode inserts, containing adaptor sequences specific for either the Illumina or SOLiD sequencing platforms, were constructed. For the Illumina-specific barcode, the barcode consisted of a 16 nt degenerate sequence that immediately followed part of the Illumina adaptor sequence, such that sequence reads would start with the barcode (Figure 4.7). This ensured that sequence reads would commence with a heterogeneous sequence, which would facilitate cluster registration and improve base-calling accuracy. The barcode region was amplified for sequencing using primers that introduced the whole of the Illumina adaptor sequences as well as a barcode for indexing samples, which was located at the end of the region of interest in the sequence read at positions 31-36 (Figure 4.9A). Thus, the required sequence data fitted within the most accurate portion of an Illumina sequence read.

For the SOLiD-specific barcode, the barcode consisted of a 15 nt degenerate sequence flanked by part of the P1-T adaptor and the internal adaptor (Figure 4.8). This ensured sequence reads on the SOLiD platform would also commence with the heterogeneous barcode sequence. The barcode region was amplified using primers that introduced the remaining portion of the P1-T adaptor sequence, as well as the P2
Chapter 4: Generation and analysis of barcoded vectors

Figure 4.7: Schematic diagram of primer extension method used to construct a barcode insert containing adaptor sequences specific for sequencing on the Illumina platform. The barcode oligonucleotide was designed to contain 32 nt of the Illumina adaptor sequence, followed by a 16 nt degenerate sequence that comprised the barcode, followed by 14 nt of a known sequence. A primer with homology to the 3'-end of the barcode oligonucleotide was annealed, generating a NsiI-compatible sticky end, and extended to produce the second strand of the insert. This product was cleaved using PstI, which recognises the double-stranded “CTGCAG” site highlighted in green, producing a NsiI-compatible sticky end at the other end of the barcode insert. The PstI-digested barcode insert was then cloned into the single NsiI site of the pEF1α.γc lentiviral plasmid construct. EF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; nt, nucleotide.
Figure 4.8: Schematic diagram of primer extension method used to construct a barcode insert containing adaptor sequences specific for sequencing on the SOLiD platform. The barcode oligonucleotide was designed to contain 23 nt of the SOLiD P1-T adaptor sequence, followed by a 16 nt degenerate sequence that comprised the barcode, followed by the 20 nt internal adaptor sequence. A primer with homology to the 3'-end of the barcode oligonucleotide was annealed, generating a NsiI-compatible sticky end, and extended to produce the second strand of the insert. This product was cleaved using PstI, which recognises the double-stranded “CTGCAG” site highlighted in green, producing a NsiI-compatible sticky end at the other end of the barcode insert. The PstI-digested barcode insert was then cloned into the single NsiI site of the EF1α-γc lentiviral plasmid construct. pEF1α, elongation factor-1-α; γc, common γ chain of the interleukin-2 receptor; nt, nucleotide.
Figure 4.9: PCR amplification to introduce adaptor sequences and sample indices to Illumina-specific and SOLiD-specific barcodes. (A) Primers used to amplify the Illumina-specific barcode introduced the adaptor sequences required for sequencing on the Illumina platform, as well as a 6 nt sample index used to identify samples for multiplexing in a single sequencing run. (B) Primers used to amplify the SOLiD-specific barcode introduced the P1-T and P2 adaptor sequences required for sequencing on the SOLiD platform, as well as a 10 nt sample index used to identify samples for multiplexing in a single sequencing run. nt, nucleotide.
For both barcoded libraries of unknown complexity, plasmid DNA was purified from over 100 single colonies and Sanger-sequenced (Section 2.3.7) for preliminary validation of library complexity and sequencing of a representative proportion of unique barcodes. This yielded 119 Illumina-specific barcodes and 100 SOLiD-specific barcodes with known and unique sequence identities.

### 4.4.4 Analyses of barcode libraries of defined complexity containing 1, 10 or 100 known barcodes

From the analyses of the 2-barcode library and the complex libraries that contained the initial barcode design, it was clear that sequencing error limited the reliability of NGS analyses of complex barcode libraries. Uncertainty remained as to whether there may be an analysable degree of complexity if an optimised barcode design were used. Therefore, barcode libraries of defined complexity containing known sequences were prepared, in order to define the limit of analysable library complexity by testing whether NGS analysis could resolve low to moderate complexity barcode libraries.

#### 4.4.4.1 Preparation of barcode libraries of defined complexity

Barcode libraries of defined complexity comprising known sequence identities were produced by mixing the plasmids containing the known barcodes in equimolar ratios. For the Illumina-specific barcode, plasmids containing known barcode sequences were mixed to provide libraries containing 10 known and 100 known barcode sequences, the “10-barcode” and “100-barcode” libraries, for NGS. A single barcode was also sequenced using NGS. To produce the SOLiD-specific 100-barcode library, 100
plasmids containing unique and defined barcode sequences were mixed in equimolar ratio with the exception of 4 plasmids mixed in 0.5× equimolar ratio, 4 plasmids mixed in 2× equimolar ratio and 2 plasmids mixed in 4× equimolar ratio for the purpose of analysing barcodes present at different abundances. Subsequent analyses of the SOLiD-specific library corrected for this difference.

4.4.4.2 Sequencing of Illumina-specific barcode libraries of defined complexity

The barcode regions were PCR amplified from the single barcode, the 10-barcode and 100-barcode libraries, as well as the highly complex barcoded library of unknown complexity. These samples were sequenced with 141,878,276 reads generated. Initial filtering based on the presence of a defined sequence at positions 17-30 yielded approximately 100,000-fold coverage of each of the known barcodes in the defined libraries (Table 4.6). The 100-barcode library was also sequenced using a second run of the HiSeq2000, yielding 825,086-fold coverage of each of the expected sequences.

Elimination of sequence reads with Phred scores below 30 filtered out 16.8%, 13.7% and 16.2% of reads for the 1-barcode, 10-barcode and 100-barcode libraries, respectively, and 9.0% of reads of the 100-barcode library for the second sequencing run (Table 4.6). The number of unique barcode sequences in each sample was subsequently counted, with and without clustering with 1 mismatch.

4.4.4.3 Analyses of the Illumina-specific 1-barcode sample

For the sample containing a single Illumina-specific barcode, 92.7% of sequence reads were accounted for by the expected sequence after Q30 filtering and clustering with 1 mismatch (Table 4.6). The remaining 7.3% of reads comprised 8,099 putatively false
Table 4.6: Effect of analytical strategies on the reduction of background caused by sequencing error for the Illumina-specific barcode libraries of defined complexity

<table>
<thead>
<tr>
<th>Property</th>
<th>1-barcode sample (Run 1)</th>
<th>10-barcode library (Run 1)</th>
<th>100-barcode library (Run 1)</th>
<th>100-barcode library (Run 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold coverage of each barcode</td>
<td>134,364×</td>
<td>81,714×</td>
<td>211,057×</td>
<td>825,086×</td>
</tr>
<tr>
<td>Counting unique barcode sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of processed sequence reads</td>
<td>134,364</td>
<td>817,135</td>
<td>21,105,664</td>
<td>82,508,636</td>
</tr>
<tr>
<td>Total unique barcodes detected</td>
<td>9,691</td>
<td>1,073</td>
<td>7,977</td>
<td>14,672</td>
</tr>
<tr>
<td>Number of processed reads accounted for by expected sequence(s)</td>
<td>122,882</td>
<td>800,981</td>
<td>20,596,796</td>
<td>80,639,047</td>
</tr>
<tr>
<td>Proportion of processed reads accounted for by expected sequence(s)</td>
<td>91.45%</td>
<td>98.02%</td>
<td>97.59%</td>
<td>97.73%</td>
</tr>
<tr>
<td>Proportion of background i.e. processed reads not accounted for by expected sequence(s)</td>
<td>8.55%</td>
<td>1.98%</td>
<td>2.41%</td>
<td>2.27%</td>
</tr>
<tr>
<td>Counting unique barcode sequences after filtering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of processed sequence reads</td>
<td>111,766</td>
<td>705,473</td>
<td>17,694,272</td>
<td>75,076,957</td>
</tr>
<tr>
<td>Total unique barcodes detected</td>
<td>8,219</td>
<td>819</td>
<td>3,638</td>
<td>5,241</td>
</tr>
<tr>
<td>Number of processed reads accounted for by expected sequence(s) after Q30 filtering</td>
<td>103,365</td>
<td>702,313</td>
<td>17,573,274</td>
<td>74,638,541</td>
</tr>
<tr>
<td>Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering</td>
<td>92.48%</td>
<td>99.55%</td>
<td>99.32%</td>
<td>99.42%</td>
</tr>
</tbody>
</table>
### Counting unique barcode sequences after filtering and clustering with 1 mismatch

<table>
<thead>
<tr>
<th>Counting unique barcode sequences after filtering and clustering with 1 mismatch</th>
<th>Number of processed sequence reads</th>
<th>Total unique barcodes detected</th>
<th>Number of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering</th>
<th>Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering with 1 mismatch</th>
<th>Proportion of background after Q30 filtering and clustering with 1 mismatch i.e. processed reads not accounted for by expected sequence(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of processed sequence reads</td>
<td>111,766</td>
<td>8,100</td>
<td>103,559</td>
<td>7.52%</td>
</tr>
<tr>
<td></td>
<td>Total unique barcodes detected</td>
<td>705,473</td>
<td>530</td>
<td>704,889</td>
<td>0.45%</td>
</tr>
<tr>
<td></td>
<td>Number of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering</td>
<td>17,694,272</td>
<td>687</td>
<td>17,636,460</td>
<td>0.68%</td>
</tr>
<tr>
<td></td>
<td>Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering with 1 mismatch</td>
<td>75,076,957</td>
<td>1,014</td>
<td>74,878,064</td>
<td>0.58%</td>
</tr>
<tr>
<td></td>
<td>Proportion of background after Q30 filtering and clustering with 1 mismatch i.e. processed reads not accounted for by expected sequence(s)</td>
<td>92.66%</td>
<td>99.92%</td>
<td>99.67%</td>
<td>99.74%</td>
</tr>
<tr>
<td></td>
<td>Proportion of background after Q30 filtering and clustering with 1 mismatch i.e. processed reads not accounted for by expected sequence(s)</td>
<td>7.34%</td>
<td>0.08%</td>
<td>0.33%</td>
<td>0.26%</td>
</tr>
</tbody>
</table>

### Counting unique barcode sequences after filtering, clustering with 1 mismatch and excluding error-prone positions

<table>
<thead>
<tr>
<th>Counting unique barcode sequences after filtering, clustering with 1 mismatch and excluding error-prone positions</th>
<th>Number of processed sequence reads</th>
<th>Total unique barcodes detected</th>
<th>Number of processed reads accounted for by expected sequence(s) after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions</th>
<th>Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions</th>
<th>Proportion of background after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions i.e. processed reads not accounted for by expected sequence(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of processed sequence reads</td>
<td>113,198</td>
<td>8,351</td>
<td>104,796</td>
<td>92.58%</td>
</tr>
<tr>
<td></td>
<td>Total unique barcodes detected</td>
<td>717,794</td>
<td>545</td>
<td>717,194</td>
<td>99.92%</td>
</tr>
<tr>
<td></td>
<td>Number of processed reads accounted for by expected sequence(s) after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions</td>
<td>18,109,912</td>
<td>706</td>
<td>18,050,957</td>
<td>99.67%</td>
</tr>
<tr>
<td></td>
<td>Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions</td>
<td>75,532,263</td>
<td>902</td>
<td>75,332,860</td>
<td>99.74%</td>
</tr>
<tr>
<td></td>
<td>Proportion of background after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions i.e. processed reads not accounted for by expected sequence(s)</td>
<td>92.58%</td>
<td>99.92%</td>
<td>99.67%</td>
<td>99.74%</td>
</tr>
<tr>
<td></td>
<td>Proportion of background after Q30 filtering, clustering with 1 mismatch and excluding error-prone positions i.e. processed reads not accounted for by expected sequence(s)</td>
<td>7.42%</td>
<td>0.08%</td>
<td>0.33%</td>
<td>0.26%</td>
</tr>
</tbody>
</table>
barcodes. Nonetheless, the distribution of the relative frequency of the 500 most abundant barcodes detected in the Illumina-specific single barcode sample indicated that the expected barcode had greater than 4 log-fold higher abundance than background (Figure 4.10A and 4.10F). This indicated that foreground had been successfully resolved from background and it would be feasible to eliminate background by applying a frequency-based cut-off.

4.4.4.4 Analysis of the Illumina-specific 10-barcode library

The number and proportion of background sequences were lower for the Illumina-specific 10-barcode library compared to the 1-barcode sample. The 10 expected barcodes accounted for 99.9% of all sequence reads of the Illumina-specific 10-barcode library after Q30 filtering and clustering with 1 mismatch (Table 4.6). Although 520 putatively false barcodes were detected, the distribution of the relative frequency of the 500 most abundant barcodes indicated the 10 expected barcodes had an approximately 3 log-fold higher abundance than background (Figure 4.10B and 4.10F). Interestingly, the 10 expected barcodes were not detected in the expected equal proportion, even though they were successfully resolved from background.

4.4.4.5 Analyses of the Illumina-specific 100-barcode library using 2 independent sequencing runs

For the Illumina-specific 100-barcode library, the number and proportion of background sequences were also low (Table 4.6), however, there was an overlap between foreground and background (Figure 4.10F). After Q30 filtering and clustering, the 100 expected barcode sequences accounted for 99.9% of sequence reads, with the remaining background including 677 putatively false barcodes. However, the
Figure 4.10: Distribution of the relative abundance of the 500 most abundant barcode sequences detected following analysis of the defined barcode libraries using different sequencing platforms. Libraries containing (A) 1 (B) 10 and (C) 100 defined Illumina-specific barcode(s) sequenced using the first sequencing run. (D) Library containing the same 100 defined Illumina-specific barcodes sequenced using the second sequencing run. (E) Library containing 100 defined SOLiD-specific barcodes. (F) Mean and range of relative abundances of expected and putatively false barcodes, for each sample.
distribution of the relative frequency of the 500 most abundant barcode sequences displayed no clear distinction between the 100 expected barcodes and background, and certain expected barcodes were detected at lower abundance than putatively false barcodes (Figure 4.10C). The first 89 most abundant barcodes matched expected sequences, and a point of inflection in the distribution of the relative frequencies of barcodes occurred at the 82nd-most abundant barcode. Six putatively false barcodes were detected within the top 100, which were not expected in the Illumina-specific 100-barcode library. Three of the expected barcodes were detected outside of the top 100, and 3 of the expected barcodes were not detected at all.

When the 100-barcode library was sequenced using 4.2-fold higher fold-coverage in an independent run, a similar pattern was observed. The 100 expected barcodes accounted for 99.7% of all sequence reads after Q30 filtering and clustering, and 914 putatively false barcodes were detected (Table 4.6). Although baseline levels of background were lower compared to the previous sequencing run of the same sample, the distribution of the 500 most abundant barcodes displayed a similar overlap between the 100 expected barcodes and background (Figure 4.10D and 4.10F). Seven putatively false barcodes were detected within the top 100. A point of inflection occurred at the 79th-most abundant barcode and again, the first 89 most abundant barcodes matched expected sequences. Five of the expected barcodes were detected outside of the top 100, and 2 of the expected sequences were undetected. Therefore, a barcode library of moderate complexity containing 100 known sequences could not be resolved from background produced by sequencing error.
4.4.4.6 Analysis of putatively false barcodes detected in the Illumina-specific 100-barcode library

The high frequency false positive barcodes were likely to represent erroneous reads of the expected sequences. Interestingly, the 6 putatively false barcodes detected in the top 100 during the first sequencing run were also detected in the top 100 during the second sequencing run. The 2 most abundant of these putatively false barcodes could not be detected by qPCR in the original plasmid mixture containing the 100 expected barcodes, using primers designed to specifically amplify those sequences (Table 4.7; Figure 4.11). This demonstrated that these sequences were false positives produced by sequencing error. The other 4 putatively false barcodes detected in the top 100 in both sequencing runs also failed to be detected by non-quantitative PCR, and similarly represent false positives produced by sequencing error.

To examine the possibility that clustering with 1 mismatch was insufficient to eliminate high frequency false barcodes, each of the false barcodes within the 120 most abundant barcodes detected in both sequencing runs were compared to the 100 expected sequences. Each of the false barcodes differed by a minimum of 6 nt from the expected sequence to which they were most closely related, indicating these high frequency false barcodes were substantially scrambled. In addition, 6 mismatches exceeded the maximum of 5 mismatches that could be tolerated in order for the 100 expected sequences to be uniquely identifiable. Therefore, further clustering would not eliminate the high frequency false positive sequences.

4.4.4.7 Analysis of the SOLiD-specific 100-barcode library

Having failed to detect each of the expected barcodes at the expected frequency and in a higher abundance than background for the Illumina-specific 100-barcode library, a
Table 4.7: Relative abundance of specific barcode sequences within the 100-barcode library as measured by the Illumina HiSeq2000 and qPCR.

<table>
<thead>
<tr>
<th>Barcode sequence</th>
<th>Rank (first sequencing run)</th>
<th>Rank (second sequencing run)</th>
<th>Relative abundance (first sequencing run)</th>
<th>Relative abundance (second sequencing run)</th>
<th>Relative abundance (qPCR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATTATTACAGGGAC</td>
<td>1</td>
<td>2</td>
<td>2.59%</td>
<td>2.78%</td>
<td>0.84%</td>
</tr>
<tr>
<td>TCGAGCTCTCGTCGC</td>
<td>2</td>
<td>3</td>
<td>2.55%</td>
<td>2.71%</td>
<td>2.43%</td>
</tr>
<tr>
<td>ATTCTTTGCAGTTAT</td>
<td>3</td>
<td>1</td>
<td>2.45%</td>
<td>2.99%</td>
<td>1.47%</td>
</tr>
<tr>
<td>TGTTAGAAAAACGAAA</td>
<td>49</td>
<td>85</td>
<td>0.93%</td>
<td>0.27%</td>
<td>0.94%</td>
</tr>
<tr>
<td>GTACGCCTCTCGTG</td>
<td>50</td>
<td>32</td>
<td>0.92%</td>
<td>1.26%</td>
<td>1.31%</td>
</tr>
<tr>
<td>ACGACCCACAGGGG</td>
<td>51</td>
<td>51</td>
<td>0.91%</td>
<td>0.89%</td>
<td>3.27%</td>
</tr>
<tr>
<td>AGCTAGCCGGGTGG</td>
<td>n.d.</td>
<td>159 (4028)</td>
<td>0.00%</td>
<td>0.000012%</td>
<td>1.46%</td>
</tr>
<tr>
<td>ACTCTGGACTTGTT</td>
<td>n.d.</td>
<td>n.d.</td>
<td>0.00%</td>
<td>0.00%</td>
<td>4.10%</td>
</tr>
<tr>
<td>GTGTGTGTGTGTGTCT</td>
<td>73</td>
<td>64</td>
<td>0.68%</td>
<td>0.71%</td>
<td>6.85%</td>
</tr>
<tr>
<td>CCTCACAAGCCAC</td>
<td>n.d.</td>
<td>n.d.</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.12%</td>
</tr>
<tr>
<td>CTTCAATGACTGTT</td>
<td>90</td>
<td>90</td>
<td>0.17%</td>
<td>0.14%</td>
<td>0.00%</td>
</tr>
<tr>
<td>GGCTTACTCTCTAG</td>
<td>91</td>
<td>93</td>
<td>0.08%</td>
<td>0.05%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*a* Relative abundance of barcodes as calculated after Q30 filtering and clustering with 1 mismatch. qPCR, quantitative PCR.

*b* n.d., not detected.

*c* Rank for the first and sequencing runs before Q30 filtering and clustering with 1 mismatch.

*d* False barcodes generated by sequencing error.
Figure 4.11: qPCR confirmation of the absence of false barcodes produced by sequencing error in the Illumina-specific 100-barcode library. qPCR standard curves for (A) the first-most abundant false barcode and (B) the second-most abundant false barcode (blue gradient), generated using monoclonal plasmids into which those false barcodes were inserted. These false barcode sequences were neither detected in the original plasmid mixture containing the 100 expected barcodes (red), nor in 3 different monoclonal plasmids (yellow, orange and brown). The 2 false barcode sequences were detected in a mixture of 10 plasmids that included those false barcodes (green). qPCR, quantitative PCR.
100-barcode library was sequenced using the SOLiD sequencing platform, to test whether an independent technology was more suitable for analysing complex barcode libraries.

The SOLiD-specific 100-barcode library was sequenced with 12,191-fold coverage of each barcode. After filtering based on the presence of 10 adaptor sequences and clustering with 1 mismatch, the 100 expected barcodes accounted for 97.33% of all sequence reads and 6,863 putatively false barcodes were detected (Table 4.8). The distribution of the 500 most abundant sequences indicated that the expected sequences could not be distinguished from background (Figure 4.10E and 4.10F). The first 82 most abundant barcodes matched expected sequences, however, 13 false barcodes were detected in the top 100. Eleven of the expected barcodes were detected outside of the top 100 and 2 of the expected barcodes were not detected. Therefore, the SOLiD platform was similarly unable to resolve a moderately complex barcode library.

4.4.5 Validation of the Illumina-specific 100-barcode library

4.4.5.1 Discrepancy between the actual and detected abundance of barcodes

Analyses of the Illumina-specific 10-barcode and 100-barcode libraries had revealed a discrepancy between the expected equal abundance of each of the known barcode sequences and the detected abundance of those sequences. For the 10-barcode library, in which each barcode was expected at 10% abundance, there was a 3.7-fold discrepancy between the most frequently detected known barcode, detected at 17.3% abundance, and the least frequently detected known barcode, detected at 4.7% abundance.
Table 4.8: Effect of analytical strategies on the reduction of background caused by sequencing error for the SOLiD-specific barcode libraries of defined complexity

<table>
<thead>
<tr>
<th>Effect of analytical strategy</th>
<th>Fold coverage of each barcode</th>
<th>12,191×</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting unique barcode sequences</td>
<td>Number of processed sequence reads</td>
<td>1,219,079</td>
</tr>
<tr>
<td>Total unique barcodes detected</td>
<td>112,624</td>
<td></td>
</tr>
<tr>
<td>Number of processed reads accounted for by expected sequence(s)</td>
<td>680,938</td>
<td></td>
</tr>
<tr>
<td>Proportion of processed reads accounted for by expected sequence(s)</td>
<td>55.86%</td>
<td></td>
</tr>
<tr>
<td>Proportion of background i.e. processed reads not accounted for by expected sequence(s)</td>
<td>44.14%</td>
<td></td>
</tr>
</tbody>
</table>

| Counting unique barcode sequences after filtering | Number of processed sequence reads | 546,409 |
| Total unique barcodes detected | 7,068 |
| Number of processed reads accounted for by expected sequence(s) after Q30 filtering | 529,264 |
| Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering | 96.86% |
| Proportion of background after Q30 filtering i.e. processed reads not accounted for by expected sequence(s) | 3.14% |

| Counting unique barcode sequences after filtering and clustering with 1 mismatch | Number of processed sequence reads | 546,409 |
| Total unique barcodes detected | 6,963 |
| Number of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering | 531,689 |
| Proportion of processed reads accounted for by expected sequence(s) after Q30 filtering and clustering with 1 mismatch | 97.33% |
| Proportion of background after Q30 filtering and clustering with 1 mismatch i.e. processed reads not accounted for by expected sequence(s) | 2.67% |
The observed, uneven distribution of abundances for each of the known barcodes in the 100-barcode library, expected at 1% abundance, followed a pattern that was consistent overall but non-identical for both sequencing runs (Figure 4.12A). The most frequently detected known barcodes were detected at 2.6% and 3.0% abundance in both sequencing runs (Table 4.7). The 3 barcodes detected at the highest abundance in the first sequencing run were reproducibly detected at the highest abundance in the second sequencing run. The observed abundance of the remaining barcodes gradually decreased, with the 89th-most abundant barcodes detected at 0.21% and 0.15% abundance by the first and second sequencing runs, respectively. Fourteen and 18 barcodes were detected at below 0.5% abundance and 3 and 2 barcodes were undetected by the first and second sequencing runs, respectively. The sequence identities of the undetected barcodes overlapped between the 2 sequencing runs. One of the undetected barcodes, “AGCTAGTCCGGGTG”, was detected at low abundance in the unfiltered and unclustered data from both sequencing runs. Empirical validation of the 100-barcode library was therefore performed to confirm the presence of and measure the relative abundance of barcode sequences.

The presence and abundance of specific barcodes in the 100-barcode library were validated using qPCR assays designed to quantify 10 specific barcodes within the original plasmid mixture of 100 known barcodes that had served as a pre-amplification template for the sequencing amplicons (Figure 4.13). The limit of sensitivity of this qPCR approach for detecting specific barcodes was established as at least 0.1%, based on qPCR analysis of a barcode sequence that was absent from the 100-barcode library and was spiked into the 100-barcode library in known proportions. The specificity of amplification of that barcode sequence was validated by Sanger-sequencing.
Figure 4.12: Analysis of the relative abundance, GC content and likelihood of secondary structure formation for each of the 100 expected Illumina-compatible barcode sequences. (A) Relative abundance of the 100 expected Illumina-specific barcode sequences, as detected during the first and second sequencing runs. (B) Distribution of the relative abundance of each barcode sequence as a function of the percentage GC content of that sequence. (C) Distribution of the relative abundance of each barcode sequence as a function of the Minimum Gibbs Free Energy (MFE) value calculated for that sequence. MFE values provide an estimate of the likelihood of secondary structure formation, with lower values associated with a higher likelihood.
Figure 4.13. qPCR confirmation of the presence of selected expected barcodes in the Illumina-specific 100-barcode library. qPCR standard curves (blue gradient) for (A) the third-most and (B) the fifty-first-most abundantly detected barcodes in the 100-barcode library, and (C) one of the undetected barcodes, generated using monoclonal plasmids containing those barcodes. These barcode sequences were detected in the original plasmid mixture that constituted the 100-barcode library (red) and in a mixture of 10 plasmids that included those barcode sequences (purple), but not in 3 different monoclonal plasmid samples (yellow, orange and pink). qPCR, quantitative PCR.
Barcodes were selected for qPCR analysis based on the abundance at which they were detected in the first sequencing run, and consisted of the 3 most abundant, the 3 mid-abundant and the 3 undetected barcodes (Table 4.7). Since the plasmid molecule containing 1 of the undetected barcodes contained a second barcode insert, which was detected by NGS analysis, that second barcode insert was also quantified. Within the limits of sensitivity of each of the qPCR assays, each of the barcodes was detected at 0.84%–6.85% abundance, including the 3 barcodes that had been undetected in both sequencing runs, which were detected at 2.12%–6.85% abundance.

To investigate a potential mechanism to account for the over-representation and under-representation of certain barcodes, the properties of the 100 barcode sequences in the Illumina-specific 100-barcode library were characterised. The percentage of GC content did not correlate with abundance, and the variance in GC content was high for both high-abundance and low-abundance barcodes (Figure 4.12B). The likelihood of secondary structure formation was predicted by calculating Minimum Gibbs Free Energy (MFE) values for each barcode amplicon. MFE values also did not correlate with barcode abundance, and the amplitude in MFE values was small overall (Figure 4.12C). Finally, multiple sequence alignments and analyses of conserved regions did not reveal any common motifs within the 10 most abundant and 10 least abundant barcodes.

4.4.6 Characterisation of empirical error frequency, error location and substitution error type for sequencing runs using the Illumina platform

Given the frequent detection of false positive sequences caused by sequencing error, the empirical error frequency, the location of errors within the barcode, and the types of substitution-like errors were analysed. Using the sample containing a single Illumina-
specific barcode, the average empirical error frequency across the barcode positions was calculated as 4.6% ± 0.2% (standard deviation) after Q30 filtering and 5.6% ± 0.3% (standard deviation) before Q30 filtering. However, since the proportion of background detected for that sample was higher than that for the Illumina-specific 10-barcode and 100-barcode libraries, it is possible that an error frequency calculated using the sample containing 1 barcode may over-represent the true error frequency. Empirical error frequency analysis was not possible using the Illumina-specific 10-barcode and 100-barcode libraries because those samples contained multiple reference sequences.

The effect of clustering with 1 mismatch upon the reduction of background for the Illumina-specific 1-barcode sample and 10- and 100-barcode libraries indicated that most sequencing errors were single nucleotide substitution-like errors, as opposed to insertion or deletion of nucleotides. Therefore, the 1-mismatch errors were characterised after Q30 filtering for the 1-barcode sample and the 10- and 100-barcode libraries analysed in the first sequencing run, and the 100-barcode library analysed in the second sequencing run. Analyses revealed that the locations of errors were not evenly distributed and differed between sequencing runs (Figure 4.14A). Errors at positions 2 or 9 accounted for 56.9% of 1-mismatch errors that arose during the first sequencing run. This differed significantly from an expected even distribution. By contrast, 41.7% of 1-mismatch errors during the second sequencing run occurred at the sixth position, which also differed significantly from an expected even distribution. Elimination of the more error-prone positions resulted in fewer reads being discarded, but overall levels of background were unchanged (Table 4.6).
Figure 4.14: Analysis of the position and substitution-like type of error for all 1-mismatch sequence errors for both Illumina HiSeq2000 sequencing runs. One-mismatch errors were compared to the known barcode sequences from which they were derived. Errors from the first sequencing run represent the sum of 1-mismatch errors for the 1-barcode, 10-barcode and 100-barcode libraries, although 1-mismatch errors from the 100-barcode library comprise 95.3% of all errors in that analysis. Errors from the second sequencing run represent 1-mismatch errors for the 100-barcode library. (A) Distribution of 1-mismatch errors across each position of the barcode (positions 2-16 of the sequence reads). This distribution differed significantly from an expected even distribution ($\chi^2 = 30,064$, df = 14, p < 0.0001 for the first sequencing run; $\chi^2 = 90,717$, df = 14, p < 0.0001 for the second sequencing run) (B) Distribution of each possible substitution-like error type. This distribution also differed significantly from an expected even distribution ($\chi^2 = 26,127$, df = 11, p < 0.0001 for the first sequencing run; $\chi^2 = 82,229$, df = 11, p < 0.0001 for the second sequencing run). df, degrees of freedom.
Analyses of each of the 12 possible types of substitution-like errors indicated that these were also unevenly distributed (Figure 4.14B). There was an over-representation of substitutions to G and an under-representation of substitutions to C, both of which were reproducible across both independent sequencing runs.

### 4.4.7 Analyses of a highly complex barcode library of unknown complexity

#### 4.4.7.1 Sequencing of the highly complex Illumina-specific barcode library

The highly complex Illumina-specific barcode library was sequenced using 109,978,259 reads. After Q30 quality filtering, a total of 9,925,944 unique barcodes were detected (Table 4.9). Clustering could not be performed for this sample due to the computationally intensive nature of the clustering operation.

<table>
<thead>
<tr>
<th>Analytic strategy</th>
<th>Number of unique barcodes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting unique barcodes without quality filtering</td>
<td>12,055,550</td>
</tr>
<tr>
<td>Counting unique barcodes after Q30 quality filtering</td>
<td>9,925,944</td>
</tr>
</tbody>
</table>

#### 4.4.7.2 Serial dilution approach to analysing the complexity of the highly complex Illumina-specific barcode library

To assist with analysing the number of unique barcodes in the highly complex Illumina-specific barcode library, the complex library was serially diluted into the 100-barcode library for NGS analysis. Ten-fold serial dilutions of the highly complex library were
performed, and the number of unique barcodes detected after Q30 filtering and clustering was multiplied by the dilution factor to estimate the complexity of the undiluted library (Table 4.10). For the 10-fold, 100-fold and 1,000-fold dilutions, the estimated library complexities were similar and in the order of 600,000. It was expected that further dilutions would dilute out the complexity of the library to a level that could be resolved from background, such as 1 or 10 unique barcodes. When the complex library was diluted further than 1,000-fold, however, the estimated library complexity increased exponentially. Furthermore, the number of unique barcodes detected appeared to relate more closely to the total number of sequence reads for each dilution. Most of these unique barcodes are likely to have represented sequencing error rather than genuine barcodes. Therefore, it was not possible to resolve the diversity of the highly complex Illumina-specific barcode library using a serial dilution approach.

<table>
<thead>
<tr>
<th>Dilution factor</th>
<th>Total sequence reads</th>
<th>Number of unique barcodes&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Number of unique barcodes, multiplied by dilution factor</th>
<th>Unique barcodes as a proportion of total sequence reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>69,907</td>
<td>58,632</td>
<td>586,320</td>
<td>83.9%</td>
</tr>
<tr>
<td>100</td>
<td>7,120</td>
<td>5,895</td>
<td>589,500</td>
<td>82.8%</td>
</tr>
<tr>
<td>1,000</td>
<td>779</td>
<td>695</td>
<td>695,000</td>
<td>89.2%</td>
</tr>
<tr>
<td>10,000</td>
<td>408</td>
<td>371</td>
<td>3,710,000</td>
<td>90.9%</td>
</tr>
<tr>
<td>100,000</td>
<td>392</td>
<td>327</td>
<td>32,700,000</td>
<td>83.4%</td>
</tr>
<tr>
<td>1,000,000</td>
<td>385</td>
<td>332</td>
<td>332,000,000</td>
<td>86.2%</td>
</tr>
<tr>
<td>10,000,000</td>
<td>272</td>
<td>253</td>
<td>2,530,000,000</td>
<td>93.0%</td>
</tr>
</tbody>
</table>

<sup>a</sup>The number of unique barcodes was counted for each dilution of the complex barcode library after Q30 quality filtering and clustering with 1 mismatch.
4.5 DISCUSSION OF CHAPTER 4 RESULTS

The growth of clinical trial activity for gene therapy targeting the haematopoietic compartment has spurred the need to develop improved methods for monitoring clonal diversity and clonal size, and in particular, greater sensitivity for detecting clonal expansions. Since the sequence identities of barcodes in a barcode library of the complexity required for clinical use would be unknown, it is critical for there to be a clear cut-off between true barcodes and background.

This chapter investigated the feasibility of vector barcoding coupled with NGS analysis for potential use in analysing clonal diversity, using libraries of defined complexity comprising known barcode sequences. This is the first study to use an empirical approach for evaluating the size of a barcode library that can be analysed using NGS. This is also the first study to utilise over 100,000-fold coverage in 2 independent sequencing runs to analyse the same barcode library, and the first study to analyse known barcode libraries using independent NGS technologies. In doing so, limitations have been found with respect to the analysable degree of complexity, sensitivity and specificity, and assessable clone size. Additionally, insights into the nature of sequencing error were uncovered.

4.5.1 Barcode libraries of low complexity can be analysed using next generation sequencing but moderately complex barcode libraries cannot

This study has shown that an understanding of NGS technology is critical to optimal barcode design. For a sample containing 2 barcodes of the initial barcode design, the expected barcodes were detected at just under a log-fold higher abundance above background. With the optimal Illumina-specific barcode design, the samples containing
1 and 10 barcodes were detected at over 4 log-fold and approximately 3 log-fold higher abundance above background, respectively. However, even with an optimally configured barcode, sequencing error imposes an upper limit on the degree of complexity that can reliably be analysed using NGS. The samples containing just 1 or 10 unique barcodes indicated that barcode libraries of low complexity could readily be distinguished from background produced by error. However, this was not the case for libraries containing 100 barcodes following sequencing on both the Illumina and SOLiD platforms (albeit with lower coverage using SOLiD).

Whilst quality filtering, clustering and exclusion of error-prone positions can reduce the total levels of background to below 0.5% of total sequence reads for a library containing 100 barcodes, it was nevertheless impossible to set a cut-off between foreground and background. Therefore, the upper limit on the analysable degree of complexity lies between 10 and 100. These findings have implications for the analysis of highly complex libraries containing over 1 million combinations of barcode sequences. Given that it has previously been estimated that a barcode library with complexity of 2,500 would be sufficient to mark 90% of transduced HPCs in a clinical trial (Gerrits et al. 2010), it is unlikely that current NGS technology can be used reliably to resolve library complexities in the order of magnitude required for clinical applications.

**4.5.2 Stochastic and systematic sequencing errors give rise to high frequency false barcodes and limit the specificity of analysis of complex barcode libraries**

The sequencing errors described in this chapter were both stochastic and systematic. The impact of stochastic errors upon the analysis of a complex library can be excluded
more readily using analytical strategies, such as the clustering approach used in this chapter. If the frequency of stochastic errors for any given position in the 15 nt barcode is known to be \( P \), or can at least be estimated as \( P \), then it is possible to calculate the proportion of error-free sequence reads as \( (1 - P)^{15} \). The proportion of reads containing at least 1 error would then be \([1 - (1 - P)^{15}]\). If a stochastic model of sequencing error were assumed, then the proportion of sequence reads containing 1 error in a barcode position, \((P)^1\), is higher than the proportion of sequence reads containing 2 errors, \((P)^2\), which in turn is higher than the proportion of sequence reads containing 3 errors, \((P)^3\), and so on. Based on the encoded degeneracy in a complex barcode library, which could be as high as \(4^{15}\) combinations of barcodes, the assumption underlying the clustering strategy was that barcodes differing by 1 position from the relatively frequent barcodes were more likely to represent stochastic sequencing error than genuine barcodes. These stochastic errors could be ‘clustered’ with the barcodes from which they were derived. Systematic errors, on the other hand, have different implications and cannot be mathematically accounted for in this manner.

The same 6 false barcodes, which differed by at least 6 nt from the expected barcode sequence to which they were most closely related, were detected in the top 100 barcodes after 2 independent sequencing runs analysing the Illumina-specific 100-barcode library, and represented high frequency systematic errors. Such high frequency errors involved the same substitution-like error occurring at the same nucleotide thousands of times. For example, the most frequently detected false barcode was counted 30,529 and 105,067 times after Q30 filtering and clustering of reads generated by the first and second sequencing runs, respectively. In a gene therapy context, such high frequency false barcodes would misleadingly suggest the existence
of non-existent clones. The detection of these false positives indicates NGS analysis of complex barcode libraries has limited specificity.

4.5.3 Limited sensitivity of next generation sequencing for detecting certain barcode sequences

The analysis in this chapter has shown that NGS analysis of barcode libraries can give rise to not only false positives, but also false negatives. In 2 independent Illumina sequencing runs, the same barcodes failed to be detected, raising the possibility that genuine clones could be missed during analysis of clinical gene therapy samples. This is of particular concern when monitoring for clonal dominance and malignancy. The danger that a malignant clone could be marked with a sequence that cannot be detected or is difficult to detect limits the potential of barcoding as a tool for clinical monitoring if contemporary NGS technologies are used. Such a malignant clone may have to attain a greater level of dominance before expansion is identified as abnormal, risking further disease progression in a patient. The problem of limited sensitivity is also associated with existing methods for integration site analysis, currently the gold-standard methods for analysing clonal complexity (Section 1.5.4; Gabriel et al. 2009; Harkey et al. 2007; Wu et al. 2013). In that context, this problem can only be addressed by using multiple restriction endonucleases that recognise a range of sequence motifs.

4.5.4 Limited potential for quantitative analysis

The potential for quantitative and unbiased analysis was an attractive advantage of barcoding over integration site methodologies. For a barcode library of moderate complexity, containing 100 unique barcodes, only a crude approximation of clone size could be obtained. Furthermore, there was over 1 log-fold difference in the measured
abundance of barcodes that had been mixed in equal proportion, despite the measures taken to minimise PCR bias by using 10 amplification cycles. Following from the analyses of the sequence characteristics of the 100 expected barcodes, it remains unclear whether and how the barcode sequences themselves can contribute to discrepancies in the measured relative abundance of barcodes. If sequence characteristics associated with impaired barcode detection could be predicted, then perhaps such characteristics could be avoided during library construction.

4.5.5 A possible alternative barcode design

A barcode could be designed to mitigate the effect of sequencing error through the use of built-in features that enable error correction. Such a barcode could comprise, for example, tandem copies of known trinucleotide repeats. An error in 1 nucleotide of the repeat could be corrected by reference to the other 2 nucleotides. In practice, however, the construction of a complex barcode library with such a configuration would be difficult as it could not be produced by existing methods for synthesising oligonucleotides containing random nucleotides. Each of the individual barcode sequences would have to be designed and synthesised separately.

4.5.6 Possible implications for next generation sequencing analyses of TCR repertoire

These findings imply that sequencing error may also limit the utility of NGS for analysing highly diverse TCR repertoires. Such analyses face similar difficulties owing to the unknown number of TCR sequences of unknown identity. In a similar study to the present one, Nguyen et al. analysed 3 single TCR sequences using the Illumina platform and detected over 500 false TCR sequences for each sample (Nguyen et al.
Validation to the degree carried out in this chapter using defined barcode libraries would be required before definitive conclusions can be drawn about the reliability of NGS for identifying genuine TCR clonotypes within a more complex population. Nonetheless, there may be useful applications of TCR repertoire analysis that involve analyses of lower complexity TCR repertoires, such as monitoring for residual disease in patients treated for leukaemia, especially when the malignant clonotype is known (Logan et al. 2011; Wu et al. 2012).

4.5.7 Insights into sequencing error

This chapter has also provided important insights into the nature of sequencing error, and has revealed that there are biases towards certain types of sequencing error. Most of the sequencing errors generated on the Illumina platform in this study were single nucleotide substitution-like errors. Many of these errors occurred at specific positions in a sequencing run-dependent manner and many errors involved specific types of substitutions in a sequencing run-independent manner. Such biases in error types underscore the value of including a defined monoclonal barcode sequence, which in this study facilitated optimisation of analytical strategies, including the identification of any error-prone positions.

The observation of biases towards certain types of sequencing errors is consistent with the known literature (Kinde et al. 2011), although the specific biases observed in this study have not previously been reported. For example, the study of single TCR sequences found most erroneous reads could be accounted for by C to T, G to A, A to G and T to C substitutions (Nguyen et al. 2011). Whilst the possibility of PCR error during preparation of the barcode amplicons cannot be excluded in this
study, the effect is likely to be minimal based on the short sequence of the barcode and the use of minimal amplification, at 10 PCR cycles, and the polymerase with the highest available fidelity. Analysis of the position-wise distribution of error frequencies for the initial barcode library, which was amplified using 62 cycles, indicated error frequencies increased from approximately position 35 to the end of the sequence read. This skewed distribution is unlikely to have been produced by PCR error.

Finally, this chapter has uncovered limitations to the utility of base-call quality values, which provide a measurement of confidence in base-calling accuracy. Analyses of the predicted and empirically observed sequencing accuracy from the first sequencing run described in this chapter showed that predicted accuracy based on quality values can overestimate true accuracy by 2 log-fold. While it is possible that this overestimation is isolated to that specific sequencing run, this may have implications for other applications of NGS which may rely on quality values during data analysis for filtering higher-quality data.

4.6 CONCLUSIONS FOR CHAPTER 4

In summary, this chapter has shown that sequencing error is a major limitation to the analysis of complex barcode libraries. Even with very high coverage of a defined library of moderate complexity, containing 100 known barcodes, it was impossible to distinguish the expected barcodes from false barcodes. Sequencing error thus imposes a limitation on the degree of complexity that can be resolved using NGS. Nonetheless, there may be applications for barcoding that involve lower orders of complexity and fall within the analysable limit, such that the potential of NGS technology can be better capitalised upon.
The conclusions that can be drawn from this chapter are:

(1) A primer extension-based methodology is an efficient method for generating a complex barcoded plasmid library;

(2) Sequencing error associated with 2 independent NGS technologies limits the capacity to analyse the number and relative abundance of unique barcodes within a moderately complex barcoded plasmid library, containing 100 unique barcodes;

(3) Sequencing error associated with the Illumina platform is both stochastic and systematic; and

(4) Barcode libraries with low complexity, containing 10 different barcodes, can be analysed using NGS, however, barcode libraries with moderate levels of complexity, containing 100 different barcodes, cannot be analysed.
CHAPTER FIVE:

GENE THERAPY RESEARCHERS’ ASSESSMENTS OF RISKS AND PERCEPTIONS OF RISK ACCEPTABILITY IN CLINICAL TRIALS

The results described in this chapter have been reported in the literature: Deakin CT, Alexander IE, Hooker CA and Kerridge IH, “Gene therapy researchers’ assessments of risks and perceptions of risk acceptability in clinical trials” Molecular Therapy (2013) 21(4): 806-15 (reproduced in Appendix 3).

5.1 INTRODUCTION

While the promise of gene therapy has long been recognised, it is only in recent years that this promise has begun to be realised. Therapeutic efficacy following gene therapy has now been reported for several inherited diseases, including SCID-X1 (Cavazzana-
Chapter 5: Survey of risk assessment in gene therapy trials

Calvo et al. 2000; Hacein-Bey-Abina et al. 2010; Gaspar et al. 2004; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011), ADA-SCID (Aiuti et al. 2002; Aiuti et al. 2009; Gaspar, Cooray, Gilmour, Parsley, Zhang et al. 2011), CGD (Ott et al. 2006), LCA (Bainbridge et al. 2008; Maguire et al. 2008), ALD (Cartier et al. 2009), β-thalassemia (Cavazzana-Calvo et al. 2010), and haemophilia B (Nathwani et al. 2011); as well as leukaemia (Kalos et al. 2011; Brentjens et al. 2011) and HIV (Mitsuyasu et al. 2009; Sections 1.2.4–5 and 1.3.1). The lag in the transition from promising early laboratory results to clinical benefit underscores the challenges and uncertainties faced in the clinical translation of complex therapeutic products. Such challenges continue to pose difficulties for gene therapy research and are likely to confront other emerging research fields, including research in stem cell therapy and nanomedicines.

One of the most significant areas of translational uncertainty concerns the extent to which therapeutic efficacy and theoretically predicted risks evaluated in preclinical models will be observed in human research subjects, and the possibility that unknown side-effects may occur. The development of leukaemia in 5 out of 20 infants treated in the French and British gene therapy trials for SCID-X1 particularly demonstrates the difficulties with predicting risks in clinical research, when the only available knowledge about safety and efficacy is generated in preclinical models with limited predictive capacity (Sections 1.3.2–1.3.3 and 1.4; Hacein-Bey-Abina et al. 2003; Hacein-Bey-Abina et al. 2008; Howe et al. 2008). Although vector-mediated insertional mutagenesis had been predicted as a possible risk, the likelihood of this resulting in leukaemia was considered remote on the basis of preclinical data obtained from cell culture and small animal models (Section 1.4.1; Hacein-Bey et al. 1996; Hacein-Bey et al. 1998; Stocking et al. 1993).
Whilst preclinical studies provide quantitative read-outs of safety and efficacy in disease models, the interpretation of data during the planning and conduct of clinical trials involves extrapolation. It is well-established that different people have different perceptions of risk based, for example, on socio-demographic factors and different experiences (Section 1.7.2; Slovic 2000; Renn & Rohrmann 2000; Dosman et al. 2001; Glendon et al. 1996; Slovic 1987). This is important because interpretation of preclinical data during the planning of clinical trials may depend upon researchers’ attitudes to risk, benefit and the validity of the preclinical models used (Sections 1.7.2–1.7.3). Attitudes towards risk may become more influential when there is greater uncertainty about extrapolating from preclinical safety and efficacy testing. For example, an understanding of the limitations of model systems used to generate preclinical safety and efficacy data may lead to an appreciation that a degree of uncertainty remains with respect to possible outcomes in humans. When researchers are balancing the potential risks and benefits of a clinical trial in these kinds of circumstances, they may attach greater weight to their perceptions of the need for a clinical trial, based on the severity of the disease and the existence and limitations of treatment alternatives. Ultimately, some uncertainty about possible risks in a phase I clinical trial is inescapable, and therefore a degree of risk must be accepted (Sections 1.7.1–1.7.2). Defining acceptable levels of risk therefore requires both scientific and moral judgment.

Little is presently known about how researchers view risks in clinical research, and about how researchers’ attitudes towards risks and perceptions of acceptable risk levels can influence their assessments of risks during the design and conduct of clinical trials. This chapter describes the results of a survey that was conducted to investigate gene therapy researchers’ assessments of risks and benefits and their perceptions of
acceptable levels of risks in phase I clinical trials. The factors selected for investigation in this study as potential contributors to decision-making about clinical trials of gene therapy included the nature of the clinical context for decision-making and researchers’ attitudes towards uncertainty, SAEs and the validity and utility of preclinical models, as revealed by researchers’ interpretations of different levels of preclinical evidence. Demographic factors, such as age, gender, professional background as clinician or scientist, and cultural background, were also investigated. Researchers’ perceptions of both the clinical context and the strength of preclinical evidence exerted an influence on decisions about initiating a phase I clinical trial. Additionally, a number of subjective and experiential influences on decision-making were identified, including length and kind of professional experience, personal experience of disease and experience with SAEs.

5.2 CHAPTER 5 HYPOTHESIS AND AIMS

The hypothesis underlying the investigations in this chapter was:

(1) Gene therapy researchers’ decisions about the design and conduct of clinical trials are influenced by their perceptions of the nature of the clinical context for decision-making and their attitudes towards uncertainty, SAEs and the validity and utility of preclinical models.

The aims of this chapter were:

(1) To investigate whether researchers’ decisions about the design and conduct of a hypothetical clinical trial are influenced by the nature of the clinical context for decision-making;
(2) To investigate whether researchers’ decisions about the design and conduct of a hypothetical clinical trial are influenced by their attitudes towards uncertainty, SAEs and the validity and utility of preclinical models; and

(3) To identify possible subjective and experiential influences on researchers’ decision-making about both hypothetical and real clinical trials.

5.3 MATERIALS AND METHODS

All work described in this chapter was performed personally by the author, except where otherwise is indicated.

5.3.1 Study design

A quantitative internet-based survey was developed for investigating researchers’ assessments of risks and perceptions of acceptable risk levels in phase I clinical trials, using a combination of short vignette questions, questions about researchers’ attitudes to specific ethical issues about risks in clinical research, and questions about researchers’ personal experience of decision-making about risk assessment (Appendix 2). The vignette questions required respondents to make hypothetical decisions about whether it would be appropriate to recruit participants to a phase I clinical trial. Respondents were presented with 8 different descriptions of preclinical data (Table 5.1) in the context of 7 different clinical scenarios (Table 5.2), and thus made a choice in 56 different permutations.

This approach required respondents to assess the risks in each hypothetical vignette and also revealed respondents’ perceptions about acceptable levels of the
### Table 5.1: Descriptions of preclinical evidence provided to survey participants for decision-making about the initiation of a clinical trial

<table>
<thead>
<tr>
<th>Abbreviated description of preclinical evidence</th>
<th>Description of preclinical evidence as provided to respondents in the survey instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell culture only</td>
<td>A statistically powered study demonstrating therapeutic efficacy in relevant primary human cells in culture AND evidence of safety in a validated cell culture assay</td>
</tr>
<tr>
<td>Cell culture only, knock-out mouse available in 2 years</td>
<td>A statistically powered study demonstrating therapeutic efficacy in relevant primary human cells in culture AND evidence of safety in a validated cell culture assay, if a knock-out mouse model is expected to be generated within 2 years</td>
</tr>
<tr>
<td>Good mouse model</td>
<td>A statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no evidence of toxicity or adverse events in long-term follow up, using a mouse model where the phenotype is the same as the human disease (no large animal model available)</td>
</tr>
<tr>
<td>Poor mouse model</td>
<td>A statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no evidence of toxicity or adverse events in long-term follow up, using a mouse model that involves the same gene but where the phenotype differs from the human disease (no large animal model available)</td>
</tr>
<tr>
<td>Good mouse model, large animal study would delay trial by 3 years</td>
<td>A statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no adverse events, using a mouse model where the phenotype is the same as the human disease AND when a large animal model is available but a large animal study would delay trial initiation by 3 years</td>
</tr>
<tr>
<td>Large animal model</td>
<td>A statistically powered study demonstrating safety and therapeutic efficacy in a large animal model, in correcting a disease phenotype that is the same as the human disease and with no adverse events in long-term follow up</td>
</tr>
<tr>
<td>Good mouse model and data from related phase I trial with no adverse events</td>
<td>Convincing preclinical safety and efficacy data in a mouse model AND data from a phase I trial targeting a different disease of the same target tissue, which used the same technological intervention and demonstrated therapeutic efficacy and no adverse events</td>
</tr>
<tr>
<td>Good mouse model and data from related phase I trial with low frequency adverse events</td>
<td>Convincing preclinical safety and efficacy data in a mouse model AND data from a phase I trial targeting a different disease of the same target tissue, which used the same technological intervention and demonstrated therapeutic efficacy and a low frequency of serious and life-threatening adverse events (e.g. less than 10%)</td>
</tr>
</tbody>
</table>
Table 5.2: Clinical scenarios provided to survey participants for decision-making about the initiation of a clinical trial

<table>
<thead>
<tr>
<th>Scenario abbreviations</th>
<th>Description of clinical scenario as provided to respondents in the survey instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1</strong></td>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
</tr>
<tr>
<td>No available treatment, death in infancy</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 2</strong></td>
<td>No available treatment, rapidly progressive, poor quality of life, death in childhood/adolescence</td>
</tr>
<tr>
<td>No available treatment, death in childhood/adolescence</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 3</strong></td>
<td>No available treatment, slowly progressive, poor quality of life during final 10 years, death in early adulthood</td>
</tr>
<tr>
<td>No available treatment, death in adulthood</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 4</strong></td>
<td>No available treatment, quality of life acceptable until end-stage, life expectancy reduced by 20 years</td>
</tr>
<tr>
<td>No available treatment, life expectancy reduced by 20 years</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 5</strong></td>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. Quality of life in survivors is reduced but generally regarded as acceptable</td>
</tr>
<tr>
<td>Disease curable by bone marrow transplant, with risks and side-effects</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 6</strong></td>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and quality of life are reduced, although quality of life is considered acceptable in most cases</td>
</tr>
<tr>
<td>Disease controlled by blood transfusions</td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 7</strong></td>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and quality of life is generally acceptable</td>
</tr>
<tr>
<td>Disease controlled by diet</td>
<td></td>
</tr>
</tbody>
</table>
potential risks in each of the hypothetical scenarios. The descriptions of preclinical data in this section were designed to reveal respondents’ attitudes towards uncertainty, SAEs and the validity and utility of preclinical models. The scenarios were constructed to elucidate respondents’ perceptions of the need for a novel treatment, and incorporated a range of disease severities, life expectancies, patient ages, qualities of life, availabilities and limitations of existing treatments, and uncertainties of outcome if the existing clinical course were followed. The questions about specific ethical issues about risks were designed to show researchers’ attitudes towards patients’ needs, the strength of preclinical evidence, and the acceptability of research involving children. The questions about researchers’ personal experience of decision-making asked about the influence of potential and actual SAEs. These questions about researchers’ attitudes and experience were designed to give further insights into how researchers assess risks and perceive acceptable risk levels.

An internet-based approach for the survey was chosen to minimise the time required to complete the survey and to facilitate representation of international gene therapy researchers, even though the trade-off was likely to be a low response rate. The survey was distributed by the ASGCT, the European Society for Gene and Cell Therapy (ESGCT), the Australasian Gene Therapy Society (AGTS), the British Society for Gene and Cell Therapy (BSGCT) and the French Society of Cell and Gene Therapy (SFTCG) to their mailing lists. In addition to this mail-out, the survey was sent directly to principal investigators of phase I trials of gene therapy, whose names were identified using the clinical trial registration website ClinicalTrials.gov and whose contact details could be accessed publicly. This enabled the most directly relevant data to be obtained from researchers who had personally been involved in clinical trial decision-making.
For all survey distribution strategies, a single follow-up request was sent a week after the initial distribution of the survey.

Research was conducted with the approval of the Human Research Ethics Committee of the Children’s Hospital at Westmead (Approval Number LNR-2011-05-04, obtained 21 April 2011).

5.3.2 Construction and validation of the survey instrument

Survey questions were developed to investigate researchers’ perceptions and assessments of risks in clinical trials (Appendix 2). Survey questions were multiple choice. Part A of the survey instrument collected the demographic details of respondents. Part B of the survey contained the 56 permutations of hypothetical questions about whether it would be appropriate to recruit participants to a phase I trial, and respondents were given the options of “strongly agree”, “agree”, “disagree” and “strongly disagree”. The hypothetical clinical scenarios were designed to represent examples of real diseases which are treated in phase I trials of gene therapy. Naming of specific diseases and usage of detailed medical jargon were avoided to ensure accessibility for researchers without medical training. The numbers of scenarios and descriptions of preclinical data in Part B of the survey reflected a balance between covering a representative range of variables contributing towards the nature of the clinical scenario, and minimising the amount of time required to complete the survey. Parts C and D of the survey instrument examined researchers’ attitudes to ethical issues about risks in clinical research and SAEs, and included questions about researchers’ personal experience of decision-making about the conduct of clinical trials.
Chapter 5: Survey of risk assessment in gene therapy trials

The internet-based survey instrument was constructed using the open source online survey application LimeSurvey, hosted by the LimeService platform (LimeSurvey.com, Hamburg, Germany). The survey instrument was validated for readability and content validity by piloting on a small number of researchers from the field.

5.3.3 Data analysis and interpretation

Statistical analysis of the survey data was conducted using the Statistical Package for the Social Sciences software (SPSS; Version 20, IBM, Armonk, NY, United States) and heat-maps were generated using R (R Core Development Team 2012). Due to the limited number of respondents, related responses were combined, such as “agree” and “strongly agree”, “disagree” and “strongly disagree”, “very relevant” and “somewhat relevant”, and “completely irrelevant” and “somewhat irrelevant”. Trends across the whole survey cohort were analysed using generalised estimating equations (GEEs), a method for modelling binary outcomes and which allows for intra-individual correlations (Menard 2008; Rabe-Hesketh & Skrondal 2008; Zeger et al. 1988). GEE analysis enabled examination of patterns in each individual respondent’s willingness to support a clinical trial, averaged across the whole survey cohort. The clinical scenario, the type of preclinical model and professional experience in the clinical care of children were examined as variables in the GEE model. The effects of changing the clinical scenario and preclinical model were also examined individually, by fitting separate GEE models for each preclinical model and for each clinical scenario, respectively.

Cross-tabs analyses were performed to analyse the association of binary outcomes with different demographic groups, using Pearson χ-squared tests, Fisher’s
exact tests and Mantel-Haenszel odds ratios. In all analyses, a p-value of <0.05 was deemed statistically significant.

Respondents were aggregated on the basis of gender (female or male), age (40 years and under or over 40 years), regional location (North American or Europe), years of research experience (10 years and under, 11–20 years, or over 20 years), professional role (scientist or clinician/clinician-scientist), and involvement in the clinical care of children (yes or no/not applicable; with and without the exclusion of non-clinicians). Responses from Australia and Asia were excluded from the analysis based on regional location, to enable investigation of cultural differences in risk perceptions in the regions with the greatest volume of clinical trial activity, that is, North America and Europe.

5.4 RESULTS

5.4.1 Demographics of respondents

A total of 156 complete survey responses were received from the mail-out to 1733 members of gene therapy societies and 287 principal investigators of phase I gene therapy trials who received direct invitations to participate. This yields a response rate in the range of 7.7%–9.0%, depending upon the extent of cross-membership between these groups, which is unknown because data from respondents was de-identified.

All major demographic groups of interest were represented, including gender, age, experience in a clinical trial, regional location, years of research experience, professional role, and involvement in the clinical care of children (Table 5.3). These demographic groups were used for the subsequent cross-tabs analyses. Seventy-five per cent of respondents reported involvement in a clinical trial, almost 60% of respondents
conducted their research in North America and almost 70% of respondents identified themselves as scientists.

Table 5.3: Demographic details of survey respondents

<table>
<thead>
<tr>
<th>Demographic group</th>
<th>Proportion and number of responses for demographic categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>38.5% Female (n=60)</td>
</tr>
<tr>
<td>Age category</td>
<td>61.5% Male (n=96)</td>
</tr>
<tr>
<td></td>
<td>31.4% 40 years and under (n=49)</td>
</tr>
<tr>
<td></td>
<td>68.6% Over 40 years (n=107)</td>
</tr>
<tr>
<td>Experience in a clinical trial</td>
<td>75% Yes (n=117)</td>
</tr>
<tr>
<td>Regional location</td>
<td>25% No (n=39)</td>
</tr>
<tr>
<td></td>
<td>57.7% North America (n=90)</td>
</tr>
<tr>
<td></td>
<td>19.9% Europe (n=31)</td>
</tr>
<tr>
<td></td>
<td>22.4% Australia, Asia and other (n=35)</td>
</tr>
<tr>
<td>Professional role</td>
<td>69.9% Scientists (n=109)</td>
</tr>
<tr>
<td></td>
<td>4.5% Clinicians (n=7)</td>
</tr>
<tr>
<td></td>
<td>25.6% Clinician-scientists (n=40)</td>
</tr>
<tr>
<td>Involvement in the clinical care</td>
<td>16.0% Yes (n=25)</td>
</tr>
<tr>
<td>of children</td>
<td>69.2% No (n=108)</td>
</tr>
<tr>
<td></td>
<td>14.7% Not applicable (n=23)</td>
</tr>
<tr>
<td>Years of experience in research</td>
<td>38.5% 10 years or less (n=60)</td>
</tr>
<tr>
<td>field</td>
<td>37.8% 11-20 years (n=59)</td>
</tr>
<tr>
<td></td>
<td>23.7% Over 20 years (n=37)</td>
</tr>
</tbody>
</table>

5.4.2 Influence of perceptions of clinical context and attitudes towards preclinical evidence on assessments of risks in a hypothetical clinical trial

Overall, the proportion of respondents who agreed that it would be appropriate to recruit subjects to a phase I clinical trial depended upon both the hypothetical clinical
scenario and the strength of the preclinical evidence (Figure 5.1). The least amount of support for a trial was observed with the combination of the scenario where disease could be controlled by diet and the cell culture model when a knock-out mouse could be generated within 2 years. The greatest amount of support was observed with the scenario involving an untreatable disease with death in infancy and with a large animal model. For the scenarios where no treatments were available, support for initiating a clinical trial generally increased with disease severity, more rapid disease progression and younger patient age. Support was lower for the scenarios where treatments with varying limitations were available.

Analysis of population-averaged effects across the responses to the panel of questions using GEEs demonstrated levels of support for initiating a clinical trial were highly dependent on both the clinical scenario (p<0.001) and the preclinical model (p<0.001). There was also a significant interaction between the clinical scenario and preclinical model variables (p<0.001). For each preclinical model, the effects of changing the clinical scenario were examined relative to the scenario which invoked the greatest support, where disease was untreatable and resulted in death in infancy (Figure 5.2, Table 5.4). The likelihood of disagreement with initiating a trial progressively decreased as treatments became unavailable, life expectancy was reduced and the disease became more severe and progressive. Changing the clinical scenario to any where treatments were available had a highly significant effect, regardless of the type of preclinical model (Table 5.4). Changing the scenario to either of the 2 scenarios where no treatments were available and life expectancy extended into adulthood had a significant effect for most of the preclinical models. Overall, the effects of changing the clinical scenario from one which was untreatable and caused death in infancy were the strongest for the types of preclinical models which involved small or large animals.
### Figure 5.1: Percentage of respondents who agreed it would be appropriate to recruit subjects to a phase I clinical trial for each of the hypothetical clinical scenarios, based on data generated in each of the preclinical models.

The radii of the circle behind each of the percentage values represent the relative proportions of respondents who agreed it would be appropriate to recruit subjects to a phase I trial. AEs, adverse events.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>No available treatment, death in infancy</th>
<th>No available treatment, death in early adulthood</th>
<th>No available treatment, death in childhood/adolescence</th>
<th>No available treatment, life expectancy reduced by 20 years</th>
<th>Disease curable by bone marrow transplant, with risks and side-effects</th>
<th>Disease controlled by blood transfusions</th>
<th>Disease controlled by diet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell culture only, knock-out mouse available in 2 years</td>
<td>56.4%</td>
<td>53.2%</td>
<td>42.9%</td>
<td>28.8%</td>
<td>19.9%</td>
<td>16.0%</td>
<td>9.0%</td>
</tr>
<tr>
<td>Cell culture only</td>
<td>61.5%</td>
<td>57.1%</td>
<td>46.2%</td>
<td>33.3%</td>
<td>26.9%</td>
<td>19.9%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Poor mouse model</td>
<td>81.4%</td>
<td>81.4%</td>
<td>71.2%</td>
<td>56.4%</td>
<td>44.2%</td>
<td>33.3%</td>
<td>23.7%</td>
</tr>
<tr>
<td>Good mouse model, large animal study would delay trial by 3 years</td>
<td>86.5%</td>
<td>87.2%</td>
<td>76.9%</td>
<td>59.6%</td>
<td>50.0%</td>
<td>37.8%</td>
<td>30.1%</td>
</tr>
<tr>
<td>Good mouse model and data from related phase I trial with low frequency AEs</td>
<td>91.0%</td>
<td>91.0%</td>
<td>86.5%</td>
<td>71.8%</td>
<td>61.5%</td>
<td>41.0%</td>
<td>30.1%</td>
</tr>
<tr>
<td>Good mouse model</td>
<td>90.4%</td>
<td>90.4%</td>
<td>88.5%</td>
<td>73.7%</td>
<td>64.7%</td>
<td>54.5%</td>
<td>39.1%</td>
</tr>
<tr>
<td>Good mouse model and data from related phase I trial with no AEs</td>
<td>96.2%</td>
<td>96.8%</td>
<td>96.8%</td>
<td>93.6%</td>
<td>82.1%</td>
<td>76.9%</td>
<td>64.7%</td>
</tr>
<tr>
<td>Large animal model</td>
<td>97.4%</td>
<td>99.4%</td>
<td>98.7%</td>
<td>93.6%</td>
<td>90.4%</td>
<td>83.3%</td>
<td>72.4%</td>
</tr>
</tbody>
</table>
Figure 5.2: Consistent patterns of odds ratios when separate GEE models were fitted for each of the preclinical models. Odds ratios representing the likelihood of disagreeing with the initiation of a clinical trial when the hypothetical clinical scenario was changed from the scenario involving an untreatable disease and death in infancy, which invoked the greatest support. Separate GEE models were fitted for each of the types of preclinical model. AEs, adverse events; BMT, bone marrow transplant; freq, frequency; GEE, generalised estimating equation.
Table 5.4: Analysis of the effects of varying clinical scenarios for GEE models fitted individually for each of the types of preclinical evidence

<table>
<thead>
<tr>
<th>Preclinical evidence</th>
<th>Scenario 7</th>
<th>Scenario 6</th>
<th>Scenario 5</th>
<th>Scenario 4</th>
<th>Scenario 3</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
</tr>
<tr>
<td>Primary human cells</td>
<td>13.08</td>
<td>&lt;0.001</td>
<td>6.45</td>
<td>&lt;0.001</td>
<td>4.35</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Primary human cells, mouse model in 2 years</td>
<td>13.13</td>
<td>&lt;0.001</td>
<td>6.78</td>
<td>&lt;0.001</td>
<td>5.22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Good mouse model</td>
<td>14.64</td>
<td>&lt;0.001</td>
<td>7.85</td>
<td>&lt;0.001</td>
<td>5.12</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Poor mouse model</td>
<td>14.08</td>
<td>&lt;0.001</td>
<td>8.76</td>
<td>&lt;0.001</td>
<td>5.52</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Good mouse model, large animal study would delay trial by 3 years</td>
<td>14.91</td>
<td>&lt;0.001</td>
<td>10.57</td>
<td>&lt;0.001</td>
<td>6.43</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Large animal model</td>
<td>14.45</td>
<td>&lt;0.001</td>
<td>7.60</td>
<td>&lt;0.001</td>
<td>4.08</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Good mouse model, data from related phase I trial without adverse events</td>
<td>13.61</td>
<td>&lt;0.001</td>
<td>7.50</td>
<td>&lt;0.001</td>
<td>5.47</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Good mouse model, data from related phase I trial with low frequency adverse events</td>
<td>23.52</td>
<td>&lt;0.001</td>
<td>14.59</td>
<td>&lt;0.001</td>
<td>6.34</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

For each of the types of preclinical evidence, odds ratios are expressed as the likelihood that support for a hypothetical trial was associated with a change in clinical scenario, relative to the scenario where there was no available treatment and death occurred in infancy. GEE, generalised estimating equation.
Similarly, separate GEE models were fitted for each of the clinical scenarios in order to examine the effects of changing the preclinical model, relative to the model that generated the least support for a trial, which was the cell culture model where a knockout mouse could be generated within 2 years (Figure 5.3, Table 5.5). The likelihood of agreeing with the initiation of a trial progressively decreased as the strength of preclinical evidence decreased, from the large animal model through to the cell culture model (Figure 5.3). For all of the clinical scenarios, the effects of changing the type of preclinical model to a small or large animal model were highly significant (Table 5.5). In general, the effect of changing the type of preclinical model was strongest for the clinical scenarios where treatments were available and quality of life was considered acceptable. Taken as a whole, the GEE analyses confirmed the overall pattern that support for a clinical trial increased with stronger preclinical models, greater disease severity and reduced life expectancy and treatment availability for the condition under study.

5.4.3 Differential risk assessment by demographic groups when making decisions about initiating a hypothetical clinical trial

To explore the potential influence of demographic factors, responses were aggregated according to gender, age, geographic location, professional role, involvement in paediatric care and years of research experience. A heat-map was employed as a visualisation tool for the purpose of guiding the analytical process. The proportions of respondents in each demographic group who supported a hypothetical clinical trial in each of the clinical scenarios and for each of the preclinical models were represented using a heat-map (Figure 5.4). The pattern of colours displayed for the total cohort of respondents in the first column of the heat-map indicated that higher support for a trial
Figure 5.3: Consistent patterns of odds ratios when separate GEE models were fitted for each of the clinical scenarios. Odds ratios representing the likelihood of agreeing with the initiation of a clinical trial when the type of preclinical model was changed from a cell culture-based model when a mouse model could be generated in 2 years. Similarly, separate GEE models were fitted for each of the clinical scenarios. AEs, adverse events; BMT, bone marrow transplant; freq., frequency; GEE, generalised estimating equations.
Table 5.5: Analysis of the effects of varying preclinical evidence for GEE models fitted individually for each clinical scenario

<table>
<thead>
<tr>
<th>Clinical Scenario</th>
<th>Large animal model</th>
<th>Good mouse model with data from related phase I without adverse events</th>
<th>Good mouse model</th>
<th>Good mouse model with data from related phase I with low frequency adverse events</th>
<th>Good mouse model, large animal study would delay trial by 3 years</th>
<th>Poor mouse model</th>
<th>Primary human cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
<td>Odds ratio (95% Wald CI)</td>
<td>Wald p-value</td>
<td>Odds ratio (95% Wald CI)</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>29.37 &lt;0.001</td>
<td>19.32 &lt;0.001</td>
<td>7.27 &lt;0.001</td>
<td>7.84 &lt;0.001</td>
<td>4.97 &lt;0.001</td>
<td>3.38 &lt;0.001</td>
<td>1.24 0.004</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>136.32 &lt;0.001</td>
<td>26.55 &lt;0.001</td>
<td>8.27 &lt;0.001</td>
<td>8.92 &lt;0.001</td>
<td>5.98 &lt;0.001</td>
<td>3.85 &lt;0.001</td>
<td>1.17 0.013</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>102.31 &lt;0.001</td>
<td>40.13 &lt;0.001</td>
<td>10.19 &lt;0.001</td>
<td>8.54 &lt;0.001</td>
<td>4.43 &lt;0.001</td>
<td>3.28 &lt;0.001</td>
<td>1.14 0.129</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>36.02 &lt;0.001</td>
<td>36.02 &lt;0.001</td>
<td>6.92 &lt;0.001</td>
<td>6.28 &lt;0.001</td>
<td>3.64 &lt;0.001</td>
<td>3.19 &lt;0.001</td>
<td>1.23 0.106</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>37.90 &lt;0.001</td>
<td>18.43 &lt;0.001</td>
<td>7.40 &lt;0.001</td>
<td>6.45 &lt;0.001</td>
<td>4.03 &lt;0.001</td>
<td>3.20 &lt;0.001</td>
<td>1.49 0.002</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>26.21 &lt;0.001</td>
<td>17.46 &lt;0.001</td>
<td>6.27 &lt;0.001</td>
<td>3.64 &lt;0.001</td>
<td>3.19 &lt;0.001</td>
<td>2.62 &lt;0.001</td>
<td>1.30 0.056</td>
</tr>
<tr>
<td>Scenario 7</td>
<td>26.66 &lt;0.001</td>
<td>18.63 &lt;0.001</td>
<td>6.51 &lt;0.001</td>
<td>4.38 &lt;0.001</td>
<td>4.38 &lt;0.001</td>
<td>3.16 &lt;0.001</td>
<td>1.24 0.081</td>
</tr>
</tbody>
</table>

For each of the clinical scenarios, odds ratios are expressed as the likelihood that support for a hypothetical trial was associated with a change in type of preclinical evidence, relative to the evidence generated in a cell culture model where a mouse model could be generated within 2 years. GEE, generalised estimating equation.
Figure 5.4: Heat-map representing the proportions of respondents in each demographic group who supported initiation of a clinical trial, for each of the clinical scenarios and types of preclinical models. Dark green represents 100% agreement with initiating a trial, while dark red represents 100% disagreement (0% agreement). The columns are separated by white spaces according to the demographic groups of interest, with the total cohort represented in the left-hand-most column. Each of the 8 groups of 7 rows separated by a white space represents 1 of the 8 types of preclinical evidence, ordered according to decreasing strength from top to bottom. These 8 rows contain groups of 7 rows, which represent each of the 7 different clinical scenarios. The 7 clinical scenarios are ordered according to decreasing disease severity from top to bottom: (1) no available treatment, death in infancy; (2) no available treatment, death in childhood or adolescence; (3) no available treatment, death in early adulthood; (4) no available treatment, life expectancy reduced by 20 years; (5) disease curable by bone marrow transplantation, with risks and side-effects; (6) disease can be controlled by blood transfusions; (7) disease can be controlled by diet. AEs, adverse events.
was correlated with increased disease severity and strength of preclinical evidence. When the respondents were aggregated according to demographic factors, the heat-map patterns for the different demographic groups were similar to the pattern for the total cohort, indicating overall that levels of support were similar. Similar colour patterns were also observed between demographic categories. In general, demographic factors appeared to exert little effect when the preclinical evidence was strong and when the disease scenarios were incurable and associated with higher severity, lower quality of life and shorter life expectancy. Gender, age, regional location and years of research experience did not appear to influence decision-making outcomes. However, in some situations where the preclinical evidence was weaker and the disease scenarios were less severe, it appeared that there were some differences in levels of support according to respondents’ professional background as a clinician or scientist and involvement in paediatric care. Therefore, the effect of professional experience on decision-making was examined for the questions involving the 2 least severe disease scenarios and the 2 cell culture-based preclinical models (Table 5.6).

When the effects of professional experience were examined, clinicians and clinician-scientists were more likely than scientists to support a phase I trial on the basis of data generated in cell culture models when disease could be treated by either blood transfusion or diet. Respondents involved in the clinical care of children were similarly more likely to support a phase I trial when the preclinical evidence was weaker and the disease was less severe. These associations persisted when non-clinicians were excluded from analysis for the cell culture model when disease could be treated by blood transfusion, but not when disease could be treated by diet. On the other hand, scientists and respondents not involved in paediatric care were more likely to decide against initiating a trial when low frequency adverse events occurred in a related
Table 5.6: Effects of professional experience on decision-making in the least severe clinical scenarios when preclinical evidence involved cell culture-based models or there was a related phase I study with low frequency adverse events

<table>
<thead>
<tr>
<th>Professional experience</th>
<th>Type of Evidence</th>
<th>Clinical Scenario</th>
<th>$X^2$ p-value</th>
<th>Odds Ratio(^a)</th>
<th>Odds Ratio p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional role</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Clinicians vs. Scientists)</td>
<td>Primary human cells</td>
<td>Disease controlled by blood transfusions</td>
<td>0.017</td>
<td>2.73</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.010</td>
<td>3.94</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Primary human cells, mouse model within 2 years</td>
<td>Disease controlled by blood transfusions</td>
<td>0.009</td>
<td>3.09</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.031*</td>
<td>3.52</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>Good mouse model with data from related phase I trial, with low frequency adverse events</td>
<td>Disease controlled by blood transfusions</td>
<td>0.006</td>
<td>2.63</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.003</td>
<td>2.96</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Changed from primary human cells, mouse model within 2 years to good mouse model</td>
<td>No available treatment, death in infancy</td>
<td>0.023</td>
<td>0.40</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No available treatment, death in childhood/adolescence</td>
<td>0.005</td>
<td>0.33</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No available treatment, death in adulthood</td>
<td>0.014</td>
<td>0.41</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease curable by bone marrow transplant, with risks and side-effects</td>
<td>0.019</td>
<td>0.43</td>
<td>0.021</td>
</tr>
<tr>
<td>Involved in clinical care of children</td>
<td>Primary human cells</td>
<td>Disease controlled by blood transfusions</td>
<td>&lt;0.001</td>
<td>5.44</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------------------</td>
<td>----------------------------------------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>(Yes vs. No or not applicable)</td>
<td>Primary human cells, mouse model within 2 years</td>
<td>Disease controlled by blood transfusions</td>
<td>&lt;0.001</td>
<td>6.57</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.008</td>
<td>4.71</td>
<td>0.008</td>
</tr>
<tr>
<td>Good mouse model with data from related phase I trial, with low frequency adverse events</td>
<td>Disease controlled by blood transfusions</td>
<td>&lt;0.001</td>
<td>3.07</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.012</td>
<td>4.86</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Primary human cells</td>
<td>Disease controlled by blood transfusions</td>
<td>0.011</td>
<td>3.09</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease controlled by diet</td>
<td>0.009</td>
<td>3.09</td>
<td>0.012</td>
</tr>
<tr>
<td>Involved in clinical care of children</td>
<td>Primary human cells</td>
<td>Disease controlled by blood transfusions</td>
<td>0.019</td>
<td>4.62</td>
<td>0.024</td>
</tr>
<tr>
<td>(Yes vs. No; non-clinicians excluded)</td>
<td>Primary human cells, mouse model within 2 years</td>
<td>Disease controlled by blood transfusions</td>
<td>0.015</td>
<td>5.50</td>
<td>0.021</td>
</tr>
</tbody>
</table>

* Indicates p-value calculated using Fisher’s exact test, which was used when counts in a group were too small for the χ² test.

Odds ratios are expressed as the likelihood that support for a hypothetical trial for the given preclinical data and clinical scenario was associated with the first of the demographic subgroups specified in the “Professional experience” column, rather than the second of the demographic groups.
phase I trial. When the preclinical model was changed from a cell culture model to a
good mouse model, scientists were more likely than clinicians and clinician-scientists
to change their decision from disagree to agree in 3 of the scenarios where treatments
were unavailable and the scenario where disease could be treated by bone marrow
transplantation.

Involvement in paediatric care had a marked impact upon decision-making
when included as a variable in the GEE models fitted for the cell culture model, the cell
culture model when a mouse model could be available in 2 years, the good mouse
model when there was a related phase I study with no SAEs, and the good mouse model
when there was a related phase I study with a low frequency of SAEs (odds ratios of
7.42, 8.54, 4.06 and 5.13, respectively; p-values of 0.006, 0.003, 0.044 and 0.024,
respectively). Similarly, professional experience in paediatric care affected decision-
making outcomes when analysed as a variable in the GEE models fitted for the
scenarios where disease could be controlled by blood transfusions or by diet (odds
ratios of 9.55 and 6.50, respectively; p-values of 0.002 and 0.011, respectively). When
non-clinicians were excluded from the GEE analysis, the effect of involvement in
paediatric care remained for the GEE models fitted for the cell culture model, the cell
culture model when a mouse model could be available in 2 years, and also the scenario
where disease could be cured by bone marrow transplantation (odds ratios of 2.75, 2.87
and 2.52, respectively; p-values of 0.049, 0.041 and 0.034, respectively).

5.4.4 Attitudes towards research involving human subjects

Responses to questions about attitudes towards research involving human subjects
correlated with the responses to decisions about initiating a hypothetical clinical trial
regarding the types of clinical scenarios and preclinical model more likely to lead to support for a trial. Respondents emphasised the importance of benefit in paediatric trials and preclinical safety testing, despite recognising the limitations of small animal models. The importance of benefit in trials involving children was expressed by almost two-thirds of respondents (Table 5.7). Over two-thirds of respondents preferred recruitment of patients with advanced disease compared to patients with early stage disease.

Although over 80% of respondents considered the absence of SAEs in small animal models did not facilitate meaningful predictions about long-term safety outcomes in humans, approximately 70% thought a clinical trial should not proceed without extensive toxicological data for a severe and incurable disease (Table 5.7). On the other hand, over 70% of respondents considered it may be acceptable to proceed to a clinical trial in the context of a severe and incurable disease, when preclinical measurements of therapeutic efficacy fell short of clinical significance in a statistically powered animal study.

When asked about their own decision-making about a clinical trial, respondents considered the potential negative impact of SAEs as a relevant factor, and indicated their personal decision-making had been affected by the report of an SAE in their own field. Nearly 80% of respondents identified the potential for SAEs to have a negative effect on public support and trust or on their own field of research as relevant considerations when planning a clinical trial (Table 5.7). Almost 60% of respondents indicated their own decision-making had been affected by the report of a SAE in a trial in their field.
### Table 5.7: Percentage of respondents who agreed with statements about attitudes towards research involving human subjects

<table>
<thead>
<tr>
<th>Statement</th>
<th>Percentage of Whole Cohort Agreeing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptability of proceeding to clinical trial without extensive toxicological data, for severe disease when treatment does not exist</td>
<td>30.8%</td>
</tr>
<tr>
<td>Absence of adverse events in small animal models do not enable meaningful predictions about long-term safety in humans</td>
<td>81.4%</td>
</tr>
<tr>
<td>It is more important to have a greater prospect of benefit in trials involving children than in trials involving adults</td>
<td>64.7%</td>
</tr>
<tr>
<td>For a progressive disease, it is preferable to recruit patients with advanced disease than patients with an early disease stage</td>
<td>68.6%</td>
</tr>
<tr>
<td>In the context of severe disease when no treatment is available, it may be acceptable to proceed to trial when preclinical measurements of efficacy fall short of statistical significance (i.e. ( p \geq 0.05 ))</td>
<td>72.4%</td>
</tr>
<tr>
<td>The potential for adverse events to have a negative effect on public support and trust is a relevant consideration when planning a trial</td>
<td>79.7%</td>
</tr>
<tr>
<td>The potential for adverse events to have a negative effect on the field of research is a relevant consideration when planning a trial</td>
<td>79.1%</td>
</tr>
<tr>
<td>Whether personal decision-making about clinical trials has been affected by the report of an adverse event</td>
<td>57.1%</td>
</tr>
</tbody>
</table>
5.5 DISCUSSION OF CHAPTER 5 RESULTS

This study investigated gene therapy researchers’ perceptions of risk, risk assessment, the validity and utility of preclinical models and SAEs, as well as the potential influence of these perceptions on decision-making about gene therapy trials. These data obtained from a subset of gene therapy researchers reveal that the characteristics of the target disease and existing therapies, as well as the availability and practicalities of animal models, are highly influential when decisions are made about the design and conduct of clinical trials. Decisions about clinical trials are influenced by a number of subjective factors, such as length and kind of experience, value-judgments regarding risk and benefit, attitudes towards the strength and validity of preclinical evidence, and perceptions of the need for a novel treatment. This represents the first empirical study of how gene therapy researchers assess risks in clinical trials, the factors which influence decision-making in clinical research, and how gene therapy researchers balance these factors. Previous empirical studies of decision-making about clinical trials have evaluated the experience of research participants, with a particular focus on consent and motivations for research participation (Joffe et al. 2001; Catt et al. 2011; Jenkins et al. 2011; Truong et al. 2011; Miller et al. 2013).

5.5.1 Study limitations

There are a number of potential limitations to this study which may arise from the relatively small sample size of 156 respondents, the heterogeneity of respondents, and limited number of qualitative questions included. The low response rate (7.7%–9.0%) is consistent with the literature regarding online surveys, which describes response rates ranging from 1% to over 60% depending upon the characteristics of the target
population, the survey length, mode and strategy for delivery, and salience of the survey topic (Cook et al. 2000; Shih, 2008). Additionally, this response rate is likely to reflect both the large number of complex questions in the survey, necessary for obtaining meaningful data, and also the likely time-poor nature of the target research population, which renders any dataset of this nature difficult to obtain. The consequence of this low response rate is that these findings are exploratory in nature, have limited potential for generalisability, and need confirmation in larger studies.

To illustrate the representativeness of this sample of 156 responses with respect to the general population of 2,020 researchers who were invited to participate, it is possible to calculate the margin of error associated with estimating a proportion of 50% in the general population, which is the most conservative proportion to estimate (National Statistical Service, 2012). The margin of error associated with a sample of 156 respondents is 7.6% with a 95% confidence interval. To achieve a margin of error of 5.0% with a 95% confidence interval, 323 respondents would be required. This suggests that the accuracy of proportions estimated using this sample population would not be substantially increased if the sample size were increased by more than 2-fold.

Related to the low response rate is the possibility of an ascertainment bias in this data, which may arise if the characteristics of survey respondents differ significantly from those of the general population of gene therapy researchers. Despite these limitations to generalisability, the findings and analyses reported here nevertheless have internal validity with respect to the sub-population of gene therapy researchers who responded to the survey. Furthermore, it is also possible that the data may be influenced by a positive ascertainment bias. Seventy-five per cent of respondents reported having previous involvement in a clinical trial, which is likely to
Chapter 5: Survey of risk assessment in gene therapy trials

be a higher proportion than that of the general population of gene therapy researchers. While the low response rate necessitates careful and qualified interpretation of the data, the 156 responses obtained in this study reflect opinions in the field and are useful for identifying factors which may influence decision-making and which can be followed up in further studies.

Although the cohort of respondents were heterogeneous with respect to professional role, years of research experience, age and cultural background, this heterogeneity also enabled identification of demographic factors which may predict attitudes and perceptions, such as professional experience. It should also be noted that few of the ‘clinicians’ in this analysis identified themselves solely as clinicians, with the majority indicating a dual role as scientists and clinicians. While this probably reflects the targeting of the survey to a community of researchers, further research and larger numbers of responses would be required to further explore differences between individuals who identify solely as either scientists or clinicians.

5.5.2 Influence of the clinical context on clinical trial decision-making

Details of the clinical context for a trial, including the particular characteristics of the disease and the availability of alternative or effective therapy, were key factors in decision-making about clinical research. This was evident from the responses of participants both to hypothetical clinical trials and to the questions regarding their attitudes to research. When patients’ needs were perceived as greater and there was increased certainty of a negative outcome under the existing clinical course, then researchers were more likely to support a clinical trial, even when the preclinical evidence was relatively weak. This may be because a limited potential for benefit,
based on the limited strength of the available preclinical evidence, may nevertheless justify a trial for patients in a severe clinical situation. By contrast, when patients generally experienced an acceptable quality of life and sub-optimal treatments were available, then respondents were more divided about the appropriateness of initiating a phase I trial. In these situations, differences emerged between demographic groups, and particularly between scientists and clinicians.

The results in this chapter make it clear that consideration of the particular clinical context and the potential for benefit shaped judgments of acceptable risk during the risk-benefit calculus for phase I trials, even though the primary goal of such trials is to evaluate safety, and not therapeutic efficacy. Respondents to this survey preferred to recruit patients with advanced disease as subjects in trials involving adults, suggesting exposure to uncertain risks was considered more acceptable when research participants had ‘less to lose’, even when the state of disease progression would reduce the potential for benefit. By contrast, in the paediatric setting, respondents generally considered the risk-benefit ratio should be skewed in favour of benefit. At the same time, when the target disease was severe and incurable, it was considered acceptable to proceed to a clinical trial even when measurements of therapeutic efficacy fell short of statistical significance in a statistically powered preclinical study, such that the prospect of benefit was unlikely. In other words, if the clinical context was dire, then it was more acceptable for the risk-benefit ratio to be skewed in favour of risk.

5.5.3 Influence of the strength and utility of preclinical evidence on clinical trial decision-making

Gene therapy researchers’ support for a hypothetical clinical trial was also strongly influenced by their attitudes towards the strength and utility of evidence generated in
different preclinical models. Even though respondents placed great store in the use of preclinical models, they were clearly mindful of the limited predictive capacity and relatedness to human disease of some preclinical models, particularly cell culture-based models and small animal models. Attitudes towards the preclinical models perceived as weaker, particularly primary human cells, were also a key point of difference between scientists and clinicians, and paediatricians in particular.

Conversely, support for initiating a clinical trial was greatest when a large animal was used as a preclinical model across all clinical scenarios, with support being almost universal when treatments were non-existent. While it is self-evident that large animal models would be a preferred standard of evidence for preclinical studies, in reality such models are not readily available, are costly to maintain and impose protracted timelines. As a consequence, it is inevitable that researchers must consider the best way to communicate to prospective research participants the remaining uncertainties about safety and efficacy following preclinical studies in small animals. These may include how the limitations of models may affect attempts to extrapolate stable phenotype correction in mice (or other demonstrations of efficacy) to long-term potential benefit in humans, the potential for differences in immune responses between species, and the impact of short life-spans of small animals on assessments of long-term safety.

5.5.4 Experiential influences on clinical trial decision-making

The possibility that scientists and clinicians may assess risks in clinical research differently highlights the influence of personal experience of disease upon decision-making; in this instance, gained through a professional role in clinical care and the care
of children in particular. This finding is consistent with earlier research comparing perceptions of risk held by experts and laypersons (Fischhoff et al. 1981; Slovic 1987). It seems likely that those who work in healthcare may have a greater ‘everyday’ experience of the burden of severe and incurable disease and the impact of such diseases on quality of life. This understanding may inform researchers’ evaluations of acceptable levels of risk in research and decisions about when it is appropriate to initiate a trial, such as whether to delay a trial in order to develop a better preclinical model or an improved vector. Conversely, scientists may have greater experience in interpreting research evidence. Involvement of both scientists and clinicians, therefore, may arguably be necessary for optimally informed decisions. Likewise, discussion and decision-making in ethics committees may also be more rigorous when committee members have a heterogeneous range of personal experiences of disease.

5.5.5 Impact of serious adverse events on clinical trial decision-making

This study provides preliminary empirical support for the notion that SAEs can delay research, not from an ethical or regulatory perspective, but as a result of researchers’ own decision-making about clinical trials. The gene therapy researchers in this study were clearly concerned not only about the short-term negative consequences of SAEs for research participants and their families, but also about the longer-term negative consequences of SAEs for public support and trust and for the gene therapy field as a whole. The majority of respondents took these latter, longer-term consequences into account when planning a clinical trial. Moreover, most researchers who had been involved in a clinical trial stated that their own decision-making had been influenced by the report of an SAE in their field. These findings suggest there can be unrecognised variables which affect decision-making. As a future direction, it would be interesting to
investigate whether other fields, which have experienced SAEs but without the same degree of media attention, would be similarly influenced.

5.6 CONCLUSIONS FOR CHAPTER 5

This survey identified a complex matrix of ethical, social, experiential and evidence-related factors which may influence decisions about clinical trials, and which may be differentially evaluated by different individuals. While the results of this survey are intriguing, further research is required to confirm the demographic predictors identified. It is unclear how explicitly and transparently these factors are considered and how they are weighed during decision-making. Judgments about risk, benefit, uncertainty and the adequacy of evidence are inescapably value-laden. This suggests that what is needed is a more sophisticated discussion about how values influence the decision-making and the risk assessments of not only researchers, but also regulators, ethics committees and research participants, and how decision-making within and between these groups may differ. It also suggests that optimal decision-making requires involvement of a mixture of people with diverse experience. A list of factors which have been identified as potential influences on decision-making and specific questions that need to be asked of any clinical study may assist the evaluation of such factors and improve transparency and rigour in decision-making about clinical trials. Such questions could be used to guide researchers’ evaluation of the preclinical evidence and clinical context, and their design of studies, selection of participants and communication of risks to potential participants (Deakin et al. 2010). This research represents a first step towards identifying some of the factors which require consideration during decision-making.
Gene therapy researchers’ assessments of risks in clinical trials are influenced strongly by their perceptions of the clinical context, the characteristics and needs of the patient cohort, and the strength and validity of preclinical evidence. Other non-scientific factors, including personal experience of disease and concerns regarding the potential negative effects of SAEs, may also influence the design and conduct of clinical trials in gene therapy. This exploratory study provides important insights into how researchers make decisions about clinical trials and makes it clear that trial design is as much a moral issue as a scientific one. Recognition that not only the decisions of researchers, but also those of research participants, ethics committees and regulatory authorities, requires the exercise of moral judgment may facilitate more sophisticated thinking about the design of first-in-human trials and the means by which such trials can be effectively communicated to potential participants.

The conclusions that can be drawn from this chapter are:

(1) Researchers’ decisions about the design and conduct of a hypothetical clinical trial are influenced by both the nature of the clinical context for decision-making and their attitudes towards uncertainty, SAEs and the validity and utility of preclinical models; and

(2) Subjective and experiential factors influence researchers’ decision-making about both hypothetical and real clinical trials. These include type and length of professional experience, attitudes towards the limitations of preclinical models, perceptions of the need for a new treatment, and attitudes towards SAEs.
CHAPTER SIX:

GENERAL DISCUSSION

6.1 INTRODUCTION

Using the outcomes of the first trials of gene therapy for SCID-X1 as an exemplar for uncertainty in translational research, a combination of biological and ethical analyses of safety and efficacy in translational gene therapy have been investigated in this thesis. The French and British trials demonstrated for the first time that gene therapy using gammaretroviral vectors can provide therapeutic benefit to patients with this severe defect of the immune system, and who may otherwise face significant risks of GvHD following a mismatched BMT. The eventual development of leukaemia in 5 out of 20 patients treated in these trials, caused by the previously underestimated risk of insertional mutagenesis, pointedly illustrated the uncertainty of risk in phase I clinical
trials and the limitations of preclinical cell culture-based and animal models for predicting outcomes in humans.

This thesis contributes to the body of knowledge that has developed in response to those SAEs in the SCID-X1 trials, which includes research into insertional mutagenesis risks and the design of safer gene transfer vectors for not only SCID-X1 gene therapy, but also other diseases affecting the haematopoietic compartment. In Chapter 3 of this thesis, the safety and efficacy profile of lentiviral vectors containing different promoter-enhancer elements for transcriptional control of γc expression was evaluated. The potential for developing and analysing barcoded vectors for the purpose of monitoring the safety and efficacy of redesigned vectors is described in Chapter 4. Finally, the ways in which gene therapy researchers assess preclinical evidence and weigh uncertain risks and benefits when making decisions about phase I clinical trials were investigated and are described in Chapter 5. By combining biological and ethical analyses in this unique manner, this thesis has shown that a critical evaluation of the interplay between biology and ethics is both inherent and essential to the clinical translation of gene therapy.

6.2 SUMMARY OF OUTCOMES OF THE BIOLOGICAL ANALYSES

6.2.1 Effect of promoter-enhancer transcriptional activity on clonal complexity and proliferative stress in reconstituted lymphocytes

In Chapter 3, lentiviral vectors containing different promoter-enhancer elements for controlling γc expression were evaluated using a murine model of SCID-X1, with a view to understanding how safety and efficacy can be influenced by features of vector design, such as the selection of an appropriate internal promoter-enhancer element.
Specifically, the effects of promoter-enhancer transcriptional activity on overall levels of lymphocyte reconstitution, the degree of clonal complexity in reconstituted lymphocytes and the development of lymphoma were investigated. In comparison to the EF1α promoter-enhancer element examined, the less transcriptionally active PGK promoter-enhancer element assessed in Chapter 3 resulted in reduced reconstitution of all lymphocyte compartments, including a marked reduction of B220⁺ B cells and CD8⁺ T cells. Despite this difference in levels of reconstituted peripheral lymphocytes, no difference between the 2 vectors was observed with respect to survival and the incidence and onset of lymphomagenesis. Analysis of the degree of lymphocyte complexity between the 2 vectors, as assessed by quantification of unique vector integration sites, was inconclusive.

The absence of a demonstrated difference between the EF1α.γc and PGK.γc vectors with respect to the risk of insertional mutagenesis-independent lymphomagenesis was unexpected. At least in part, it may relate to the difficulty of making measurements of end-points such as lymphomagenesis, which involve multiple somatic mutations and uncharacterised mechanisms. The SCID-X1 mouse model system that was used in this work may not be adequately sensitive for making measurements of this nature. As an alternative, tumour-prone mouse models, such as those described in Section 1.4.3 (Montini et al. 2006; Shou et al. 2006), could offer improved sensitivity for measuring differences in lymphomagenesis between vectors containing different internal promoter-enhancer elements, particularly if the anticipated effect size is small. Nonetheless, this study highlights the importance of evaluating the adequacy of preclinical model systems for investigating the questions under consideration. This, in turn, raises the question of the best way to interpret such data when decisions about initiating a clinical trial are made.
6.2.2 Feasibility of vector barcoding and analysis of complex barcode libraries using next generation sequencing

In Chapter 4, the feasibility of generating barcoded vectors and analysing complex barcode libraries using NGS was investigated. Such an approach may be less biased than conventionally-used integration site methodologies for quantifying clonal contributions, and thus may be useful for monitoring safety and efficacy in haematopoietic gene therapy. Complex barcoded plasmid libraries were successfully generated, however, the exact complexity of those libraries could not be analysed due to the high frequency of sequencing error which resulted in a background of false barcodes. Analyses of samples containing 1, 10 and 100 known barcode sequences indicated that while the true barcodes could readily be resolved from background in libraries of low complexity, this was not possible in a moderately complex library of 100 barcodes. Furthermore, the limited quantitative potential of NGS for analysing complex libraries was uncovered using these defined samples, in which certain barcodes were detected as over-represented or under-represented, and certain barcodes were undetected.

Although it was known that NGS technologies have relatively high error rates in comparison to Sanger sequencing, the effect of sequencing error in preventing the resolution of expected barcodes from background in a moderately complex library was unexpected. This background of false barcodes could not be resolved, despite the application of analytical strategies such as quality filtering and clustering with 1 mismatch. The work in Chapter 4 demonstrates the value of empirical approaches for evaluating the limitations of new technologies. Such limitations also highlight the need
for critique during the interpretation of data, and this may apply to the interpretation of preclinical data during clinical trial decision-making.

**6.3 LIMITATIONS OF TECHNOLOGIES, METHODOLOGIES AND MODEL SYSTEMS USED FOR PERFORMING BIOLOGICAL MEASUREMENTS AND PRECLINICAL ASSESSMENTS**

**6.3.1 Limitations of next generation sequencing technology**

The scale of sequencing enabled by the recent advent of NGS technologies has given rise to numerous, previously inconceivable, sequencing applications. Such applications include the possibility of sequencing complex barcode libraries, as investigated in this thesis. The sequencing of large numbers of unique barcode sequences is an example of rare variant analysis. This is a particularly demanding application for NGS, since it requires the identification of rare variants against a background of rare variants, of which there may be millions of combinations. These difficulties are further compounded by the heterogeneity of barcode sequences and the absence of a reference sequence against which to compare variants, such as a genomic sequence used as a reference point for the identification of single nucleotide variants.

Another study has purported to investigate the feasibility and utility of barcoded vectors, using samples that represented a complex library (Lu et al. 2011). Sequence reads consisted of a known 6 nt sample index followed by an unknown 27 nt barcode. An analytical strategy was developed to reduce background produced by sequencing error, based on the frequency of background detected for the known 6 nt sample index. The distribution of the abundance of background for sample indices, which was approximately 1 log-fold lower than the expected sequences, was used to set a threshold for the analysis of unique barcodes and to estimate a false positive rate. This
approach makes 2 assumptions, which may be invalid in light of the results of Chapter 4. The first assumption is that the distribution of background sequences is the same for both the sample indices and the barcodes. This approach may be valid if sequencing error were merely stochastic, however, the results of Chapter 4 clearly demonstrate that sequencing error can be systematic. In addition, the number of sample indices used is not specified, although it is likely to be in the order of 10 and unlikely to be as high as 100. In Chapter 4, the distribution of background differed significantly for libraries with a complexity of 10 versus 100. This implies that the parameters set by Lu et al. may have been inadequate for removing the background in their highly complex library.

The second assumption used in the strategy employed by Lu et al. (2011) is that the degree of background observed for the sample indices is representative of that for the barcode sequences. Since the barcode sequences were 21 nt longer than the index sequences, it is likely that further errors would accumulate at these additional nucleotides, resulting in a higher degree of background for the barcode sequence compared to the sample index. Chapter 4 showed that sequencing accuracy and base-calling quality decrease as read length increases. Furthermore, the validity of this strategy was not demonstrated using a complex library containing known barcode sequences, such as the 100-barcode libraries that were analysed in Chapter 4. Given that Lu et al. reported 20% of reads contained an error in the sample index, which is a large proportion for the total degree of background, it is likely that their data were impugned by a high rate of errors. Therefore, it is difficult to support the conclusion of Lu et al. that over 80,000 combinations of barcode sequences were successfully identified in a highly complex library.
Other examples of rare variant analyses that are analogous to barcoding include analyses of TCR diversity and analyses of the identity and relative abundance of haplotypes, for example, within a mixed microbial population or a genetically heterogeneous tumour. The results of Chapter 4 may also have implications for these studies. In one analysis of TCR diversity, over 1 million TCR β chain sequences were identified and used to establish a lower limit on TCR repertoire (Warren et al. 2011). However, given that the same study identified thousands of unique sequences for the 13 reference J gene segments, it is difficult to exclude the likelihood that many of the unique TCR β chain sequences identified were also false. Others have reported that the TCR repertoire is more diverse than previously expected and consists mostly of low frequency variants (Klarenbeek et al. 2010). Whilst analyses of the complex barcode libraries in Chapter 4 similarly suggested that those libraries consisted mostly of low frequency variants, the characterisation of sequencing error afforded by the analyses of defined control samples indicates that many of those variants represent sequencing error. Another study, which is analogous to the work described in Chapter 4, analysed monoclonal TCR sequences (Nguyen et al. 2011). Per nucleotide error rates of 5–6% were reported and over 10,000 unique sequences were detected in those samples, which is consistent with the results in Chapter 4.

Sequencing error has arisen as a challenge in the context of analysing single nucleotide variants within a heterogeneous population. Improved analytical strategies that seek to minimise the impact of sequencing error have been reported. Similar to the quality filtering approach described in Chapter 4, in some strategies the calling of a variant is dependent upon the quality score of that variant nucleotide (Li et al. 2008). Other strategies, such as error correction, are analogous to the clustering approach that was developed in Chapter 4, in that lower frequency sequences that deviate from higher
frequency sequences by 1 nt are presumed to derive from those higher frequency sequences (Qu et al. 2009; Zagordi et al. 2010; Zagordi et al. 2011). Other analytical strategies have purported to model sequencing run-specific error rates (Wilm et al. 2012). Whilst some of these analytical strategies are claimed to be capable of identifying variants present at below 1%, or even 0.1%, of the total population, it is evident that variants present at even lower frequencies, such as 0.0001% in a complex barcode library containing 1 million barcodes, are unlikely to be identified reliably. Furthermore, none of these strategies have been validated empirically in a manner similar to the evaluation of quality filtering and clustering that is described in Chapter 4.

Ideally, variants identified by NGS should be validated using Sanger sequencing technology. In practice, however, the extent of this re-sequencing may be restricted by the limited throughput and cost of Sanger sequencing. Furthermore, the relative frequency of individual variants may be difficult to validate as this could involve either a customised nucleotide array-based approach or separate qPCR assays dedicated to each individual variant. Consequently, it is likely that the accuracy of NGS analyses of sequence identities and relative abundances has been assumed in some of these applications.

The analyses described in Chapter 4 of defined and validated barcode libraries containing known sequences contribute useful information that may help with this difficulty of validating the output of NGS. Importantly, this work uncovered the occurrence of systematic errors, which consistently led to the detection of certain false sequences at high abundance, and demonstrated that NGS may have limited potential for conducting quantitative analysis of certain sequences. Additionally, there were
biases with respect to the position and substitution types of errors made using the Illumina platform. Predictions of error rates using quality scores were not reflective of actual error rates. These findings highlight the importance of avoiding assumptions about the capabilities of new technologies such as NGS. Massive sequence coverage and overlap can offset sequencing inaccuracies when NGS is used for its intended applications, such as genome sequencing. Such high coverage is unlikely to be afforded in applications that purport to detect rare variants. Clearly, there is a need to conduct extensive validation of the output of such technologies, particularly when applied to new and challenging applications, and to question how evidence is generated.

6.3.2 Limitations of integration site analysis methodologies

The limitations of Illumina NGS analysis described in Chapter 4 may have implications for the analyses of integration sites using Illumina NGS described in Chapter 3. For example, if the integration site dataset were similarly impugned by sequencing error, it may be possible that multiple reads of a single integration site could map to different genomic loci, thus overestimating the actual number of unique integration sites. Conversely, integration sites may be missed if the sequences of those genomic loci cannot be detected using NGS. Such limitations of NGS may compound the known limitations of integration site methodologies, outlined in Section 1.5.4, which include biases introduced by the use of restriction endonucleases and PCR. Vector-genomic junction fragments may be unamplifiable if they are located over 1 kb away from the recognition site of the restriction endonuclease used, or may be unmappable if they are located less than 20 bp away from the recognition site and the intervening genomic sequence is too short to map uniquely to the genome. Integration sites may also be
unmappable because they align ambiguously or to non-sequenced portions of the genome.

Whilst these limitations reinforce the need to question how evidence is generated and what such ‘evidence’ means, it is also clear that analyses using integration site methodologies must continue to be used in spite of the known limitations. Ultimately, these methods represent the best available for investigating retroviral integration behaviour, analysing clonal complexity and monitoring patient safety in clinical trials. Such analyses are critical for any application of gene therapy using integrating vectors to target the haematopoietic compartment. Therefore, the limitations to these methods must be tolerated to the extent that they cannot be further minimised, for example, by using multiple restriction endonucleases, or otherwise improved upon.

6.3.3 Limitations of small animal models

The limitations of small animal models for analysing safety are widely acknowledged. These include discordance between outcomes in animal experiments and human trials and problems with inter-species extrapolation (LaFollette & Shanks 1996). In Chapter 5, 81.4% of survey respondents considered that the absence of adverse events in small animal models does not enable meaningful predictions about long-term safety outcomes in humans. The finding in Chapter 3 that the choice of a less transcriptionally active internal promoter-enhancer element does not increase the risk of lymphomagenesis in mice does not necessarily eliminate this possibility in humans, given that mouse cells transform more readily than human cells (Boehm et al. 2005; Hahn & Weinberg 2002).
As outlined in Chapter 1 (Section 1.4), preclinical analyses using small animal models failed to predict the development of leukaemia in patients who were treated in the French and British SCID-X1 trials (Hacein-Bey et al. 1996; Hacein-Bey et al. 1998; Hacein-Bey-Abina et al. 2003; Hacein-Bey-Abina et al. 2008; Howe et al. 2008), albeit because of inadequate follow-up time (Kohn, Sadelain, Dunbar, et al. 2003; Baum et al. 2003). Further examples of the limitations of animal models include the failure of mouse studies to predict immune responses to AAV that have occurred in gene therapy trials for haemophilia B and lipoprotein lipase deficiency (Manno et al. 2006; Mingozzi et al. 2009).

Conversely, mouse models may predict problems that will not necessarily manifest in human clinical trials. The observation of vector-mediated hepatocellular carcinoma following administration of AAV to neonatal mice caused controversy in the liver gene therapy field (Donsante et al. 2007), perhaps particularly amongst those researchers who are keen to pioneer clinical trials using AAV. Consequently, other researchers have emphasised in their studies the absence of hepatocellular carcinoma either totally or at levels above background, and have sought to distinguish their studies, for example, on the basis of the age of animals at treatment (Bell et al. 2005; Li et al. 2011; Nathwani et al. 2011). In the absence of long-term follow-up data in humans who have undergone liver-targeted gene therapy using AAV, however, the risk of AAV-mediated hepatocellular carcinoma in humans remains an open question.

Similar to the limitations of NGS technologies and integration site methodologies, the limitations of animal models call for a need to critique how evidence is generated in preclinical studies. The limited extent to which extrapolations can be made between animal models and humans means that, ultimately, a degree of
uncertainty is unavoidable when clinical trials are initiated. Safety and efficacy outcomes observed in preclinical studies are often not observed in human trials, as has been the case for many trials of gene therapy. For example, in spite of promising preclinical data generated using a mouse model (Ferrari et al. 1991), therapeutic efficacy was not observed in the first trials of gene therapy for ADA-SCID, although this was subsequently attributed to the maintenance of enzyme-replacement therapy (Bordignon et al. 1993; Bordignon et al. 1995; Blaese et al. 1995; Kohn et al. 1995). Outcomes of a human trial may also fail to be predicted by preclinical studies, such as the SAEs that occurred in the first SCID-X1 trials (Hacein-Bey-Abina et al. 2003; Hacein-Bey-Abina et al. 2008; Howe et al. 2008).

6.4 POSSIBLE TRANSLATIONAL UNCERTAINTIES ABOUT SAFETY AND EFFICACY THAT MAY REMAIN AFTER PRECLINICAL ANALYSES

Given the limitations of technologies, methodologies and model systems used in preclinical analyses, what are the possible translational uncertainties that may remain following a preclinical analysis such as that described in Chapter 3? Both the EF1α and PGK promoter-enhancer elements that were assessed can enable sufficient γc expression to restore γc-dependent signalling via the IL-2, IL-7 and IL-15 receptors in human lymphoid cells (Smyth et al. 2007). Whether and to what extent expression from the less transcriptionally active PGK promoter-enhancer may impair reconstitution in humans, particularly of the T cell compartment, is unknown. Although it is known that the clonal complexity of peripheral T cells from patients treated in the French SCID-X1 trial correlated with the dose of infused gene-modified HPCs, the effect of different heterologous regulatory elements on peripheral complexity remains unknown in a
human in vivo setting. The limitations of integration site analysis methodologies and NGS technologies would also hinder measurements of clonal abundance.

The extent to which proliferative stress may manifest in leukaemia in human trials of haematopoietic gene therapy is also unknown. To date, no cases of insertional mutagenesis-independent leukaemia have been reported from any of the clinical trials of gene therapy where patient survival and follow-up have been sufficiently long to evaluate this risk. The doses of gene-modified HPCs used in the reported trials may have been sufficiently high to protect against proliferative stress. Alternatively, the HPC doses that could lead to proliferative stress in humans may differ from those observed in mice. Mice are known to be poor models of human tumourigenesis, and are more susceptible to tumours that occur via different mechanisms (Hahn & Weinberg 2002; Rangarajan & Weinberg 2003; Boehm et al. 2005). For example, although humans undergo approximately $10^{16}$ mitoses in a lifetime compared to mice, which undergo $10^{11}$ mitoses, approximately 30% of humans develop cancer in the course of their 70–80 year lifetime, in contrast to the 30% of laboratory mice that develop cancer in the course of a 2–3 year lifetime (Rangarajan & Weinberg 2003).

Finally, translational uncertainty will inescapably remain with respect to genotoxicity. Even with the best available small animal models, insertional mutagenesis is difficult to evaluate, and certainty with respect to extrapolating to long-term safety outcomes in humans is limited. Due to the limitations of integration site methodologies and NGS technologies, uncertainty remains as to how sensitively possible SAEs can be detected at an early stage.
6.5 THE LEVEL OF CERTAINTY REQUIRED FOR DECISION-MAKING

Since preclinical analyses cannot accurately predict the safety and efficacy of gene therapy in humans, all gene therapy clinical trials involve a degree of uncertainty, and this is inescapable irrespective of the quality and quantity of preclinical data. Indeed, this uncertainty can only be reduced by conducting more clinical trials. Uncertainty can therefore be more realistically viewed as a range of qualitative values between absolute certainty and absolute uncertainty, and it is within this range that most research and medicine operates (Deakin et al. 2010).

As a consequence of such irreducible uncertainties, key questions to consider when decisions are made about initiating a trial are: firstly, what degree of certainty about likely safety and efficacy outcomes is required to justify a clinical trial following on from preclinical studies; and secondly, what further utility is served by collecting additional data? Evaluating the degree of certainty that researchers and participants require in order to make decisions is both an empirical question concerning the predictive capacities of preclinical models, and a moral question as to whether it is possible to obtain sufficient levels of certainty to justify trials in humans (Deakin et al. 2010). These assessments hinge on the clinical context, and in particular, the burden of disease and the availability and limitations of existing treatments. Such assessments also necessarily draw upon the values and perspectives of the decision-makers, and these values also deserve consideration during the process of decision-making about gene therapy trials.
6.6 SUMMARY OF OUTCOMES OF THE ETHICAL ANALYSES

In an effort to provide insight into how the values and perspectives of decision-makers can influence decisions about a clinical trial, Chapter 5 investigated factors weighed by gene therapy researchers when asked to make a decision about initiating a hypothetical phase I clinical trial. Researchers’ decision-making was heavily dependent upon both the strength of the available preclinical evidence and the nature of the clinical context for the disease under consideration. Accordingly, decisions about clinical trial design involved both scientific and moral judgment. A number of subjective influences were found, including attitudes toward the strength and validity of preclinical data, perceptions of the need for a novel treatment, personal experience of disease, length and kind of professional experience and attitudes toward SAEs. In situations where the preclinical evidence was perceived as weaker, then perceptions of the need for a novel treatment became more influential.

6.7 INFLUENCE OF CLINICAL CONTEXT ON BIOLOGICAL ANALYSES

The severity and natural history of the diseases targeted by gene therapy research, the availability and limitations of existing treatment alternatives, and the possibility of benefit are important contextual considerations when preclinical data are evaluated during clinical trial decision-making. Such considerations may justify greater levels of risk and uncertainty. For example, the results of the survey in Chapter 5 indicated that greater uncertainty and a higher level of risk are more tolerable when the disease under consideration is severe and untreatable.

Although SCID-X1 can be treated by BMT, the associated effectiveness and risks are dependent on the particular medical centre, and are limited by the availability
of compatible donors. Survival is high for the minority of patients with access to an HLA-identical sibling donor, which is only approximately 20% of Australian patients with primary immunodeficiencies (Mitchell et al. 2013). Survival is substantially reduced, at 70%–80% at 5 years post-BMT, when haploidentical parent donors or unrelated donors are used (Section 1.2.2 and Table 1.1; Mitchell et al. 2012; Buckley 2011; Gennery et al. 2010; Haddad et al. 1998). Patients with access to an HLA-identical sibling donor are thus excluded from gene therapy trials. Results from the French and British SCID-X1 trials suggest gene therapy has produced improved safety and efficacy compared with conventional allogeneic transplantation, at least for the duration of follow-up to date (Section 1.3.1 and Table 1.4; (Hacein-Bey-Abina et al. 2010; Gaspar, Cooray, Gilmour, Parsley, Adams et al. 2011). While the development of leukaemia in 5 SCID-X1 infants was unfortunate, 4 of these infants were treated successfully by chemotherapy and continue to benefit from the gene therapy administered before the onset of leukaemogenesis.

Given the complex, subjective and highly contextual nature of risk assessment in gene therapy research, it is not unexpected that different stakeholders, including researchers, clinicians and participants, may accept different levels of risk and uncertainty in any particular clinical trial. Additionally, the same stakeholders may make different decisions in different disease contexts. The potential for different decision-making outcomes is not necessarily problematic, as subjective perspectives are to be anticipated. Moreover, the pursuit of diverse ideas and attitudes toward risk and uncertainty may result in the use of different approaches to dealing with risk and uncertainty, and thus may impart further knowledge about gene therapy interventions. For example, although the cases of leukaemia in the SCID-X1 trials were linked directly to LTR promoter-enhancer elements within the gammaretroviral vectors used
(Section 1.3.3), the Italian trial of gene therapy for ADA-SCID has continued to use a similarly configured gammaretroviral vector without incident (Aiuti et al. 2009). The safety and efficacy data obtained from the ongoing ADA-SCID trial will provide a useful point of comparison for subsequent trials of gene therapy that target the bone marrow compartment using differently configured vectors and regulatory elements. Similarly, the international SCID-X1 trial using a SIN gammaretroviral vector and the trial at St Jude Children’s Research Hospital using a SIN lentiviral vector will enable useful comparisons between SIN gammaretroviral and lentiviral vectors once sufficient patient cohorts have been established and long-term follow-up made possible (Sections 1.3.2 and 1.6.4).

Finally, since patients experience the burden of disease and are confronted by the limitations of existing treatments, it is reasonable that their experience, attitudes and values should inform evaluations by researchers about the extent to which disease context legitimately affects their judgments regarding the levels of risk and uncertainty that are acceptable in gene therapy clinical trials. Empirical studies provide some evidence of the insights that may be gained from attention to patients’ perspectives. Patients with a range of conditions, including sickle-cell anaemia, Crohn’s disease, rheumatoid arthritis and multiple sclerosis, are willing to accept more risk in return for the possibility of greater benefit (Johnson et al. 2009; Johnson et al. 2007; O’Brien et al. 1990; van Besien et al. 2001). Patients with limited treatment options may also be more tolerant of risk, even when the anticipated benefit is in alleviating symptoms or improving quality of life, rather than achieving a cure. In taking patients’ perspectives into account, it is also important to recognise that decisions made by patients to participate in trials are influenced by a range of factors, including hope for benefit, altruism, fear of uncertainty, trust of physicians and desire to access the latest
treatments (Madsen et al. 2000; Madsen et al. 2002; Nurgat et al. 2005; Stryker et al. 2006;). Greater interaction between gene therapy researchers, patients and their representative organisations is, therefore, to be encouraged.

6.8 INTERDEPENDENCE BETWEEN BIOLOGICAL AND ETHICAL ANALYSES IN THE INTERPRETATION OF RESULTS DESCRIBED IN THIS THESIS

Further insights relevant to the interpretation of each results chapter of this thesis are enabled when this body of work is considered as a whole, and when both the biological and ethical analyses are drawn upon. The biological analyses in Chapters 3 and 4 have implications for one another. As discussed in Section 6.3.1, the errors associated with NGS technologies described in Chapter 4 have implications for application of NGS to integration site analysis methods. The limitations of methodologies for integration site analysis (Sections 1.5.4 and 6.3.2) indicate that even the best available methods for analysing clonal complexity, which are currently used for monitoring clonal diversity and dynamics, have drawbacks.

The biological analyses in Chapters 3 and 4 also have implications for the interpretation of the results of the survey described in Chapter 5. For example, even though the less transcriptionally active PGK promoter-enhancer element was not associated with an increased risk of lymphomagenesis in mice, the potential remains for an association between promoter-enhancer transcriptional activity and lymphomagenesis in humans, given the known differences between murine and human tumourigenesis (Section 6.4). The results in Chapter 4 suggest that difficulties may arise when new technologies such as NGS are used to estimate risk when the limitations of such technologies are poorly understood. The 454 platform of NGS has
been used to analyse integration sites in clinical samples obtained from the French SCID-X1 trial and a trial of gene therapy for HIV-1 (Wang et al. 2010; Wang et al. 2009), however, the potential impact of sequencing error upon such analysis remains unknown. A better understanding of the limitations associated with NGS technologies, and the implications for risk knowledge generated using such technologies, may facilitate risk assessment when clinical trials are planned.

Conversely, the ethical analysis in Chapter 5 has implications for the interpretation of the biological analyses in Chapters 3 and 4. Given that preclinical assessments of risk are imperfect predictors of outcomes in human clinical trials, Chapter 5 showed that preclinical data can be evaluated differentially and subjectively by individual decision-makers. Accordingly, even though proliferative stress is not well-characterised in the context of gene therapy and the preclinical analysis described in Chapter 3 showed that vector design does not affect this risk, different researchers may interpret these data differently with respect to how much certainty about the risk of proliferative stress is required for clinical trial decision-making. The limitations of NGS technology are not yet fully understood and the analysis in Chapter 4 makes it clear that caution is required when NGS data are interpreted. There may, nevertheless, be some useful clinical applications of NGS, so long as the data output is interpreted with care and appropriate validation strategies are in place. For example, NGS technology could allow mutations to be identified in patients’ samples on a scale that would be impossible using Sanger sequencing. Candidate mutations identified using NGS would then be validated using some other method, such as Sanger sequencing.

The history of medical experimentation is one of applying new treatments, techniques and technologies in the clinic, often before these have been fully
characterised. Whilst insertional mutagenesis was evidently not well understood when
the first SCID-X1 trials were initiated, a cohort of patients have undeniably benefited
from the vectors used in those trials, even though those vectors are now considered sub-
optimal technology for gene transfer. Researchers whose experience has been informed
by the clinical need for new treatments may be less concerned about how well
characterised new technologies are when deciding to apply them to the clinic, if such
technologies have the potential to fill a pressing need.

6.9 CONCLUSIONS AND SIGNIFICANCE OF THIS WORK

The translation of promising preclinical studies from the laboratory to human
application is a challenging area of research. Even with the best configured
experimental designs using the best available model systems, a degree of uncertainty in
extrapolating from those systems to predicting outcomes in humans is unavoidable. In
addition, the methodologies and technologies used to make biological measurements in
preclinical studies have limitations that may impact upon how well the read-outs can
answer the questions under investigation. Nonetheless, and consistent with the findings
in Chapter 5, the imperative to trial novel and potentially promising interventions such
as gene therapy may be compelling, given the dire nature of the disease conditions, and
may therefore still be justifiable. Even with a vector that has a strong tendency to cause
genotoxicity, the remarkable therapeutic benefit to patients in the French and British
SCID-X1 trials is beyond dispute.

This thesis has described a unique, interdisciplinary approach to dealing with
uncertainty in the translation of gene therapy to the clinic. A critical approach to both
the biological and ethical analyses of issues that are raised in this translational realm,
and an understanding that both disciplines can usefully contribute to each other, will foster the progress of gene therapy to the clinic.
REFERENCES


APPENDIX 1: CUSTOMISED PERL SCRIPTS USED TO ANALYSE NEXT GENERATION SEQUENCING DATA IN CHAPTER 4

barcode-filter-initial.pl

Written by Mr Jeffrey Deakin

#!/usr/bin/perl
# barcode-filter -q=n (does implicit -extract).
# Will run 4x faster as files are 25% of original size and also quality values are pre-calculated.
# Can be used with data from any experiment.
# e.g.
# perl barcode-filter.pl -q=35 < Clonal_sequence.csv > seq.clonalq35
# mysqlimport -udeakin -pdeakin --local --delete sequences
# seq.clonalq35
# then

# either a) save the counts (output two columns in the correct order, so that dominant sequences appear at top)
# to allow you to later view the information implicit in the counts - perhaps visualise using gnuplot?
# mysql -ucmri -pcmri sequences < count.sql > seq.clonalq35.count
# and view or gnuplot seq.clonalq35.count

# or b) output one columns (again, in the correct order so that dominant sequences appear at top)
# the output would be the input to the clustering so seq.clonalq30.count is not actually needed
# mysql -ucmri -pcmri sequences < count.sql | awk '{print $1}' > seq.clonalq35.strings

use Getopt::Long;
GetOptions("quality=s" => 
QualityFilter); # print "qualityFilter=$QualityFilter";
while (<STDIN>) {
  @string = split /,/, $_;
  $seq = @string[0];
  $qualityString = @string[1];
  @qual = split / /, $qualityString;
  $goodQuality = 1;
  # create quality array and check for threshold quality value
  # if any element is < qualityFilter, discard the row
  for ($i = 0; $i <= $#qual; $i++) {
    if ($qual[$i] < $QualityFilter) {
      $goodQuality = 0;
    }
  }
  if ($goodQuality == 1) {
    print $_;
  }
}
cluster-initial.pl

Written by Mr Jeffrey Deakin

# Cluster algorithm to cluster 'like' sequences using Hamming method with distance <= 1

# The input to the algorithm is the output of the MySQL select
# select string as sequence, count(*) as count from seq group by 1
# order by 2 desc;
# The MySQL select statement ensures that the dominant counts appear first

# @data is the array of strings (sequences) in the file
# read data into array and remove \n chars
chomp(@data=<STDIN>);
<size> = @data;
print "records=\$size\n";
# match is a hashmap of strings, indexed by the current string, that is used to ensure a sequence
# is not processed more than once, and is populated with an arbitrary 'y'
my %match;
# list is a hashmap of arrays, indexed by the current string, and the array consists of the set
# of other strings that are 'similar' to the current string
my %list;
for ($i = 0; $i < $size; $i++) {
$input = @data[$i];
@fields = split /\t/,$input;
$thisSequence = @fields[0];
$thisSequenceCount = @fields[1];
# ignore MySQL header & do not process more than once
if ($thisSequence ne "sequence" && not exists $match{$thisSequence}) {
$list{$thisSequence} = [];
# add new hash list
for ($j = $i + 1; $j < $size; $j++) {
$input = @data[$j];
@fields = split /\t/,$input;
$mutantString = @fields[0];
if (not exists $match{$mutantString}) {
# do not process more than once
if (hd($thisSequence, $mutantString) <= 1) {
# add 'similar' mutant if <= 1 position difference
push @{$list{$thisSequence}}, $input;
# add jth elt to ith list
$match{$mutantString} = 'y';
# set jth elt to be matched
}
}
#$i % 1000 == 0 {
#$elements = $list{$thisSequence};
#$array = @$elements;
# print "finished processing record $i $thisSequence
$thisSequenceCount list=@array\n";
#
}
# for each sequence, add the MySQL counts for each mutant in the list
# to the MySQL count for the sequence,
# and output the results in the correct order of dominant sequences
print "finished processing, now add the MySQL counts\n";
for ($i = 0; $i < $size; $i++) {
  $input = @data[$i];
  @fields = split /\t/, $input;
  $thisSequence = @fields[0];
  $thisSequenceCount = @fields[1];
  if ($thisSequence ne "sequence" && exists $list{$thisSequence}) {
    print "adding counts to record $i $thisSequence\n";
    $elements = $list{$thisSequence};
    @array = @$elements;
    $arraySize = @array;
    if ($arraySize == 1) {
      $mutantString = "'similar' mutant ";
    } elsif ($arraySize > 1) {
      $mutantString = "'similar' mutants ";
    }
    for ($j = 0; $j < $arraySize; $j++) {
      $mutant = @array[$j];
      @mutantFields = split /\t/, $mutant;
      $mutantString .= @mutantFields[0];
      $mutantString .= " ";
      $thisSequenceCount += @mutantFields[1]; # add
      mutant count to sequence count
    }
    print "$thisSequence	$thisSequenceCount\n";
    if (@array > 0) { 
      print "\tincluding $mutantCount $mutantString\n";
    } else { 
      print "\n";
    }
  } else {
  }
#
# ref http://www.perlmonks.org/?node_id=500235
sub hd{ length( $_[ 0 ] ) - ( ( $_[ 0 ] ^ $_[ 1 ] ) =~ tr\[\0\][\0\] ) }

barcode-parse.pl

Written by Mr Jeffrey Deakin

#!/usr/bin/perl

# barcode-parse (input=array of 12 39,40,44,45,...) and output a
generic csv file with (seq,quality) on the same line.
# Split by !. Calc quality values and encode with " " separator
e.g. 95 90 88 92 99 98 89 etc.
# Subsequent programs will split this by " ".
# This could be reused with different input for different
experiments. This generic output file is
# independent of the experiment and is the starting point of all
subsequent analysis.
# If there are no reverse sequences, omit the -r and -o parameters
# e.g.
# barcode-parse -p=39,40,44,45,49,50,54,55,59,60,64,65 [-
f=66,17,GTTGTAACACGAAGGGC -r=37,17,GCCCTTCG7GTTACAAC -o=15] \ 
# Clonal_sequence.data > Clonal_sequence.csv
# the parameter p starts at zero ie position 1 in the string is "0"
# etc
# the parameters in [] are optional and if omitted, a single
# forward sequence is assumed
# call subroutine to get parameters from command line
&getArgs();
# &printArgs();  # debugging
$linecount = 1;  # debugging
while (<STDIN>) {
    @string = split /!/, $_;  # get current record (="$_")
    and split into @string array
    # i.e. string[0] is sequence (line 2 of original file) and string[1]
    is quality (line 4 of original file)
    $sequence = @string[0];
    $quality = @string[1];
    # use subroutine wildcard to extract the 12 wildcard sequence letters
    $seq = &wildcard("sequence", $sequence);
    # save $seq for print after reading
    quality record
    # use subroutine wildcard to extract the 12 wildcard quality values
    $qual = &wildcard("quality", $quality);
    # print "sequence $linecount $direction @$seq\n";
    # dereference
    # print "quality $linecount $direction @$qual\n";
    # dereference
    $linecount++;
    print @$seq; print ",@$qual\n";
}
# TODO use p, f & r options to parse (ie not hard copy...)
# ref http://www.sthomas.net/roberts-perl-
tutorial.htm/ch20/Returning_arrays for technique used to pass array
back and access using @$
sub wildcard {
    local($type, $string, @array);
    # define local variables
    ($type, $string) = ($_[0], $_[1]);
    # and assign values
    # if forward sequence not supplied (this is the default) or if it
    is supplied and the forward sequence is matched (as per strategy 1)
    if ($#f == 0 || ($#f >= 0 && substr($string,$fpos,length $fstr)
    eq $fstr)) {
        $offset = 0;
        $direction = "fwd";
    } else {
        $offset = $o;
        $direction = "rev";
    }
    for ($i = 0; $i <= $#p; $i++) {
        $thisPos = $p[$i];
        print "string $string array[$i]=substring($thisPos+$offset,1)\n";
        $array[$i] = substr($string,$thisPos+$offset,1);
    }
    # create quality array
}

# ref http://maq.sourceforge.net/fastq.shtml "Solexa/Illumina Read Format" for value 64 below
if ($type eq "quality") {
    for ($i = 0; $i <= $#array; $i++) {
        $temp = $array[$i];
        $array[$i] = ord($temp) - 64;
    }
}

    # wildcard: $type $direction array: @array
    return (@$array);
    # return reference to array (sequence or quality)
}

sub getArgs {
    use Getopt::Long;
    # get program arguments from command line and store in global variables
    # e.g. find "p=anything" and put anything into string $pos
    # i.e. $pos contains "39,40,44,45,49,50,54,55,59,60,64,65"
    GetOptions ("p=s" => \$pos,"f=s" => \$fwd, "r=s" => \$rev, "o=s" => \$o);
    # split $pos into array @p using "," separator
    # i.e. @p = 39 40 44 45 49 50 54 55 59 60 64 65
    # i.e. @p[0]=39 @p[1]=40 ... @p[11]=64 @p[12]=65
    # length = @p is $#p = 11 = "highest index number", total number of elements is 11+1=12 since first index is 0
    @p = split /,/, $pos;
    @f = split /,/, $fwd;
    @r = split /,/, $rev;
    # since @f = 66 GTTGTAACACGAAGGGC
    # then @f[0]=66 and @f[1]=GTTGTAACACGAAGGGC (these are strings!)
    $fpos = @f[0];
    $fstr = @f[1];
    $rpos = @r[0];
    $rstr = @r[1];
    # hard coded length=15!
    # print "size of p=$#p p=@p"
    if ($#p == 0) {
        print "Please specify p parameter e.g. -p=39,40,44,45,49,50,54,55,59,60,64,65\n";
        die;
    }
    if ($#f >= 0) {
        # print "fpos=$fpos fstr=$fstr"
    }
    if ($#r >= 0) {
        # print "rpos=$rpos rstr=$rstr"
    } else {
        # print "Reverse not specified, not parsing for reverse strings\n"
    }
}

sub printArgs {
    print "pos=$pos fwd=$fwd rev=$rev o=$o\n";
    print "p=@p f=@f r=@r o=@o\n";
    print "size of p=$#p\n";
    print "size of f=$#f\n";
    print "size of r=$#r\n";
}

290
barcode-filter.pl

Written by Mr Jeffrey Deakin

#!/usr/bin/perl
# barcode-filter -q=n (does implicit -extract).
# Will run 4x faster as files are 25% of original size and also
# quality values are pre-calculated.
# Can be used with data from any experiment.
# e.g.
# perl barcode-filter.pl -q=35 < Clonal_sequence.csv >
# seq.clonalq35
# mysqlimport -udeakin -pdeakin --local --delete sequences
# seq.clonalq35
# then

# either a) save the counts (output two columns in the correct
# order, so that dominant sequences appear at top)
# to allow you to later view the information implicit in the counts
# - perhaps visualise using gnuplot?
#   mysql -ucmri -pcmri sequences < count.sql >
# seq.clonalq35.count
# and view or gnuplot seq.clonalq35.count

# or b) output one columns (again, in the correct order so that
# dominant sequences appear at top)
# the output would be the input to the clustering so
# seq.clonalq30.count is not actually needed
# mysql -ucmri -pcmri sequences < count.sql | awk {print $1}
# > seq.clonalq35.strings

use Getopt::Long;
GetOptions("quality=s" => \$qualityFilter);
# print "qualityFilter=$qualityFilter";
while (<STDIN>) {
  @string = split /,/, $_;
  $seq = @string[0];
  $qualityString = @string[1];
  @qual = split / /, $qualityString;
  $goodQuality = 1;
  # create quality array and check for threshold quality value
  # if any element is < qualityFilter, discard the row
  for ($i = 0; $i <= $#qual; $i++) {
    if ($qual[$i] < $qualityFilter) {
      $goodQuality = 0;
    }
  }
  if ($goodQuality == 1) {
    print $_;
  }
}

cluster.pl

Written by Mr Jeffrey Deakin

# Cluster algorithm to cluster 'like' sequences using Hamming method
# with distance <= 1
# The input to the algorithm is the output of the MySQL select
# select string as sequence,count(*) as count from seq group by 1
# order by 2 desc;
# The MySQL select statement ensures that the dominant counts appear
# first
# $distance is the hamming distance
use Getopt::Long;
GetOptions ("distance=s" => \$distance);
if ($distance==0) {
  die "distance not specified\n";
}

# print "mismatches=$distance\n";
# @data is the array of strings (sequences) in the file
# read data into array and remove \n chars
chomp(@data=<STDIN>);
$size = @data;

# print "records=$size\n";
# match is a hashmap of strings, indexed by the current string, that
# is used to ensure a sequence
# is not processed more than once, and is populated with an arbitrary
# 'y'
my %match;
# list is a hashmap of arrays, indexed by the current string, and the
# array consists of the set
# of other strings that are 'similar' to the current string
my %list;
for ($i = 0; $i < $size; $i++) {
  $input = @data[$i];
  @fields = split /	/, $input;
  $thisSequence = @fields[0];
  $thisSequenceCount = @fields[1];
  # ignore MySQL header & do not process more than once
  # if ($thisSequence ne "sequence" && not exists
  # $match{$thisSequence})
  # do not process more than once
  if (hd($thisSequence, $mutantString) <= $distance) { # add 'similar' mutant
    push ( @{$list{$thisSequence}}, $input);
    # add $jth elt to $ith list
    $match{$mutantString} = 'y';
    # set $jth elt to be matched
  }
}

# if ($i % 1000 == 0) {
  # $elements = $list{$thisSequence};
  # @array = @$elements;
  # print "finished processing record $i $thisSequence
  # $thisSequenceCount list=@array\n";
  # }
# and output the results in the correct order of dominant sequences
# print "finished processing, now add the MySQL counts
"
for ($i = 0; $i < $size; $i++) {
    $input = @data[$i];
    @fields = split /\t/, $input;
    $thisSequence = @fields[0];
    $thisSequenceCount = @fields[1];
    if ($thisSequence ne "sequence" && exists $list{$thisSequence}) {
        # print "adding counts to record $i $thisSequence
        $thisSequenceCount
";
        $elements = $list{$thisSequence};
        @array = @$elements;
        $arraySize = @array;
        if ($arraySize == 1) {
            $mutantString = "'similar' mutant ";
        } elsif ($arraySize > 1) {
            $mutantString = "'similar' mutants ";
        }
        for ($j = 0; $j < $arraySize; $j++) {
            $mutant = @array[$j];
            @mutantFields = split /\t/, $mutant;
            $mutantString .= @mutantFields[0];
            $mutantString .= " ";
            $thisSequenceCount += @mutantFields[1]; # add
            mutant count to sequence count
        }
        print "$thisSequence	$thisSequenceCount\n";
        if (@array > 0) {
            print "\nincluding $mutantCount $mutantString\n";
        } else {
            print "\n";
        }
    }
}
# ref http://www.perlmonks.org/?node_id=500235
sub hd { length( $_[0] ) - ( ( $_[0] ^ $_[1] ) =~ tr\[\0]\[\0\] ) }
if (@known[$k] ne @allsequence[$k]) {    # found a mismatch
    $pos = $k + 1;
    $substitution = "@known[$k] -> @allsequence[$k]";
    # $substitutionCount{$substitution} is a hashmap of counts for
    substitution xx
    # to subsequently get hash value, $subCount =
    $substitutionCount{$substitution};
    $substitutionCount{$substitution} += 1;
    # @positionCount is an array of counts for positions 1-15
    @positionCount[$pos] += 1;
}

print "Substitution|tCount|tPercent\n";
$total = 0;
for my $key ( keys %substitutionCount ) {  
    $total += $substitutionCount{$key};
}
for my $key ( keys %substitutionCount ) {  
    $count = $substitutionCount{$key};
    $percent = sprintf "%.2f", 100 * $count / $total;
    print "$key|t$count|t$percent\n";
}
print "\n\nPosition|tCount|tPercent\n";
$total = 0;
for ($k = 0; $k < @known; $k++) {  
    $pos = $k + 1;
    $total += @positionCount[$pos]
}
for ($k = 0; $k < @known; $k++) {  
    $percent = @positionCount[$pos] / 134364;
    print "$pos|t@positionCount[$pos]|t$percent\n";
}

error-analysis.pl

Written by Mr Jeffrey Deakin

# Analysis of mutants - mismatched letters
# to run: perl ../error-analysis.pl -r=n -- cluster-file count-file
# to run in debug mode: perl ../error-analysis.pl -d=1 -r=n --
# cluster-file count-file
# n number of records in the cluster file to read (e.g. read the first
# 1 for 1-barcode, the first 10 for 10-barcode, the first 97? for 100-
# barcodes)
# cluster-file = output of clustering algorithm
# count-file = output of the MySQL count (select string as
# sequence,count(*) as count from seq group by 1 order by 2 desc;)
# (The MySQL select statement ensures that the
# dominant counts appear first)
# read data into array and remove \n chars
use Statistics::Distributions;
use Getopt::Long;
use mysql;
GetOptions ("d=i"=> \$debug, "r=i" => \$records, "count=s" => \$count);
if ($records==0) {
    die "records not specified\n";
}  
# MySQL CONFIG VARIABLES
$host = "localhost";
$database = "sequences";
$tablename = "barcode";
$user = "claire";
$pw = "claire";
# DATA SOURCE NAME
$dsn = "dbi:mysql:$database:localhost:3306";
# PERL DBI CONNECT
$connect = DBI->connect($dsn, $user, $pw);
# open(CSVFILE, @ARGV[ $#ARGV ]);
# list is a hashmap of lists used to cache the space-separated quality values, and is indexed by the current expected sequence string
# instead use barcode
# desc barcode;
# +----------+--------------+------+-----+---------+-------+
# | Field | Type | Null | Key | Default | Extra |
# +----------+--------------+------+-----+---------+-------+
# | sequence | varchar(255) | YES | | NULL | |
# | qv | varchar(255) | YES | | NULL | |
# +----------+--------------+------+-----+---------+-------+
# my %list;
# while (<CSVFILE>) {
#     chomp($_); # remove newline chars
#     @string = split /,/, $_;
#     $seq = @string[0];
#     $qualityString = @string[1];
#     instead of building hashmap of lists, insert into barcode
#     values($seq, $qualityString)
#         # later select qualityStrings from barcode where seq=sequence =>
#         populate list of quality strings (see @array below)
#         push(@{ $list{$seq} }, $qualityString);
# }
# close(CSVFILE);
open(COUNTFILE, @ARGV[ $#ARGV ]);
chomp(@countData=<COUNTFILE>);
close(COUNTFILE);
$size = @countData;
# process counts and cache them by storing the count for each sequence
# into a hashmap, indexed by the sequence
for ($i = 0; $i < $size; $i++) {
    $input = @countData[$i];
    @fields = split /	/, $input;
    $thisSequence = @fields[0];
    $thisSequenceCount = @fields[1];
    # to subsequently get hash value, $thisSequenceCount =
    $seqCount{$thisSequence} = $thisSequenceCount;
    $seqCount($thisSequence) = $thisSequenceCount;
}
open(CLUSTERFILE, @ARGV[ $#ARGV-1 ]);
chomp(@clusterData=<CLUSTERFILE>);
close(CLUSTERFILE);
$records++;
# count mutant totals
$total = 0;
for ($i = 1; $i < $records; $i++) {
    $input = @clusterData[$i-1];
    @fields = split /	/, $input;
    $mismatchString = @fields[2];
    @mismatchFields = split / /, $mismatchString;
    $size = @mismatchFields;
    if ($size==0) {
        die "record $i does not contain any mismatches\n";
    }
    for ($j = 4; $j < $size; $j++) {
        $mm = @mismatchFields[$j];
        $total += $seqCount{$mm};
    }
}
@positionCount = ();
# process mutants from cluster file
if ($debug == 1) {
    print "| Expected | Variant | Position | Substitution | Same as previous base? | tFrequency | tRelative frequency of variant | tMutant qv < Seq qv | tSequence | t(ss,n,mean) | t|tt|tMin|tMedian|tMax|
";
} else {
    print "| Expected | Variant | Position | Substitution | Same as previous base? | tFrequency | tRelative frequency of variant | tMutant qv < Seq qv | tSequence | t|t
";
}
for ($i = 1; $i < $records; $i++) {
    if ($debug == 1) {
        print "| sequence | position | t(ss,n,mean) | t\n";
    }
    $input = @clusterData[$i-1];
    @fields = split /	/, $input;
    $sequence = @fields[0];
    # lookup space-separated quality string for current sequence
    from database
    # $elements = $list{$sequence}; # old way
    # @array = @$elements;
    # @array is a list of strings containing space-separated quality
    values
    $query = "select qv from barcode where sequence='$sequence'";
    $query_handle = $connect->prepare($query);
    # EXECUTE THE QUERY
    $query_handle->execute()
    or die "SQL Error: $DBI::errstr\n";
    @array = ();
    while ($qv = $query_handle ->fetchrow_array) {
        push(@array, $qv);
    }
    $sequenceCount = @array;  # count of
    quality strings for the current sequence
    $sequenceLength = length $sequence;  # sequence string
    length
    # calculate & cache the expected statistics (sum, sum of
    squares) for each position in two arrays, indexed by quality position
    @sumx = ();
    @sumxsq = ();}
for ($k = 0; $k < $sequenceLength; $k++) {  
# for each sequence position k
$sum = 0;
$sumsq = 0;
for ($m = 0; $m < $sequenceCount; $m++) {  # process all arrays of quality values from original csv file
$qualityString = @array[$m];
@qual = split / /, $qualityString;
# quality array
$qv = @qual[$k];
# kth position
$sum += $qv;
$sumsq += $qv * $qv;
}  
$avg = sprintf "%.1f", $sum/$sequenceCount;
$pos = $k + 1;
if ($debug == 1) {
print "$sequence	$pos	($sumsq,$sequenceCount,$avg)
";
}  
push (@sumx, $sum);
push (@sumxsq, $sumsq);
}
# parse the output of the cluster program: @mismatchFields[4] are the mutant sequences
@seq = split(//,"$sequence");
$count = @fields[1];
$mismatchString = @fields[2];
@mismatchFields = split / /, $mismatchString;
$size = @mismatchFields;
if ($size==0) {
  die "record $i does not contain any mismatches
";
}
print "Start $sequence"
for ($j = 4; $j < $size; $j++) {
  $mm = @mismatchFields[$j];
  print "$sequence\t$mm"
  my @mutant = split(//,"$mm");
  for ($k = 0; $k < @seq; $k++) {
    # the kth position is where the mismatch occurs
    if (@mutant[$k] ne @seq[$k]) {
      $pos = $k + 1;
      $substitution = "$seq[$k] -> @mutant[$k]"
      print "$pos\t$substitution"
      # to subsequently get hash value,
      $subCount = $substitutionCount{$substitution};
      $substitutionCount{$substitution} += $seqCount{$mm};
      # @positionCount is an array of counts
      @positionCount[$pos] += $seqCount{$mm};
      if ($k > 0 && @mutant[$k] eq @mutant[$k-1]) {
        print "yes"
      } else {
        print "\t"
      }
  }
}
# 13/9/2012 zero-divide check

(encountered in BGI2 100-barcode data)

```
$percent = "%\%";
if ($total > 0) {
    $percent = sprintf "%.2f", 100 *
    $seqCount{$mm} / $total;
}
```

```
print "$seqCount{$mm}\t$percent\%\t";
```

lookup space-separated quality string

```
for mutant from hashmap
    # $elements = $list{$mm};
    # @array = @$elements;
    # @array is a list of strings containing
    space-separated quality values
    $query = "select qv from barcode where
    sequence='$mm'";
    $query_handle = $connect-
    >prepare($query);
    # EXECUTE THE QUERY
    $query_handle->execute()
    or die "SQL Error: $DBI::errstr\n";
    @array = ();
    while ($qv = $query_handle -
    >fetchrow_array) {
        push(@array, $qv);
    }
```

```
$nmut = @array;
@x = ();
for ($m = 0; $m < $nmut; $m++) {
    $qualityString = @array[$m];
    @qual = split / /, $qualityString;
    $qv = @qual[$k];
    push (@x, $qv);
}
```

```
$sum = 0; # sum of qvs for
```

```
mutant sequence
```

```
$sumsq = 0; # sum of squares of
```

```
qvs for mutant sequence
```

```
for ($m = 0; $m < $nmut; $m++) {
    $sum += @x[$m];
    $sumsq += @x[$m] * @x[$m]
}
```

```
$seqsumx = @sumx[$k]; # cached
```

```
sum of qvs for current sequence
```

```
$seqss = @sumxsq[$k]; # cached
```

```
sum of squares qvs for current sequence
```

```
$nseq = $sequenceCount;
# cached n is $sequenceCount
```

```
$seqMean = sprintf "%.1f", $seqsumx /
```

```
$sequenceCount;
```

```
$mean = sprintf "%.1f", $sum / $nmut;
```

```
@data = sort { $a <=> $b } @x;
```

```
$med = @data % 2 ? @data[(($nmut-1)/2) : ((@data[$nmut/2-1]+@data[$nmut/2])/2);
```

```
$median = sprintf "%.1f", $med;
```

```
$min = @data[0];
```

```
$max = @data[$nmut-1];
```

```
$pooledss = ($seqss - $nseq * $seqMean *
```

```
$nseqMean + $sumsq - $nmul * $mean) / ($nseq + $nmul - 2);
```

298
\[ s_{xy} = \sqrt{ \frac{\text{pooled} \cdot (n_{seq} + n_{mut})}{n_{seq} \cdot n_{mut}} }; \]

\[ t = \frac{n_{seq} \cdot (\text{mean} - \text{seqMean})}{s_{xy}}; \]

\[ t_{stat} = \text{sprintf "%.2f", t}; \]

\[ df = n_{seq} + n_{mut} - 2; \]

\[ t_{crit} = \text{Statistics::Distributions::tdistr}(df, 0.05); \]

\[ t_{critstat} = \text{sprintf "%.2f", t_{crit}}; \]

\[ \text{if ($sum / n_{mut} < seqsumx / sequenceCount && t > t_{crit})} { \]

\[ \text{print "$t_{critstat} (df)**\t"; } \]

\[ \text{else} \{ \]

\[ \text{print "\t"; } \]

\[ \text{if ($debug == 1) } { \]

\[ \text{print "($seqss,n_{seq},seqMean)\t($sumsq,n_{mut},mean)\t$t_{stat}\t$min\t$median\t$max
"; } \]

\[ \text{else} \{ \]

\[ \text{print "$seqMean$t_{mean}\t$t_{stat}\n"; } \]

\[ } \]

\[ } \]

\[ } # $substitutionCount{xx} is a hashmap of counts for substitution xx \]

\[ } # to subsequently get hash value, \]

\[ $subCount = \text{substitutionCount}{$substitution}; \]

\[ $seqCount{$mm}; \]

\[ @positionCount is an array of counts \]

\[ for positions 1-15 \]

\[ $seqCount{$mm}; \]

\[ print "Substitution\tCount\tPercent\n"; \]

\[ $total = 0; \]

\[ for my $key ( keys %substitutionCount ) \{ \]

\[ $total += $substitutionCount{$key}; \]

\[ } \]

\[ for my $key ( keys %substitutionCount ) \{ \]

\[ $count = $substitutionCount{$key}; \]

\[ $percent = \text{sprintf "%.2f", 100 * $count / $total}; \]

\[ print "$key\t$count\t$percent\n"; \]

\[ } \]

\[ print "Position\tCount\tPercent\n"; \]

\[ $total = 0; \]

\[ for ($k = 0; $k < @seq; $k++) \{ \]

\[ $pos = $k + 1; \]

\[ $total += @positionCount{$pos} \]

\[ } \]

\[ for ($k = 0; $k < @seq; $k++) \{ \]

\[ $pos = $k + 1; \]

\[ $percent = \text{sprintf "%.2f", 100 * @positionCount{$pos} / $total}; \]

\[ print "$pos\@positionCount{$pos}\t$percent\n"; \]

\[ } \]
gate-string.pl

Written by the author

#!/usr/bin/perl
$/ = undef;
use Getopt::Long;
GetOptions("f|filename=s"=>\$file);

open (IN, $file) or die "Error opening $file $!
";
$sequencelist = <IN>;

@sequenceList = split(/\n/, $sequencelist);

foreach $sequence(@sequenceList) {
    $barcode = chomp($sequence);
    #prints "0" instead of "$barcode", not sure why!
    $a = ($sequence =~ tr/A//);
    $t = ($sequence =~ tr/T//);
    $g = ($sequence =~ tr/G//);
    $c = ($sequence =~ tr/C//);

    $total = ($a + $t + $g + $c);
    $Apercent = ($a/$total);
    $Tpercent = ($t/$total);
    $Gpercent = ($g/$total);
    $Cpercent = ($c/$total);
    $GCpercent = (($g+$c)/$total);
    print "$barcode	"; #printf for print function, % specifies the field
    printf "A :\t %.4f\t", $Apercent; #printf for print
    printf "T :\t %.4f\t", $Tpercent;
    printf "G :\t %.4f\t", $Gpercent;
    printf "C :\t %.4f\t", $Cpercent;
    printf "GC content =\t %.4f\n", $GCpercent;
}

100-noise-hamming-distance.pl

# perl ../../../100-noise-hamming-distance.pl 100expected.csv false-in-top-120.csv > 100-noise-hamming-distance.csv
$expectedfile = @ARGV[0];
$noisefile = @ARGV[1];
open(EXPECTEDFILE, $expectedfile);
chomp(@expectedBarcode=<EXPECTEDFILE>);
open(NOISEFILE, $noisefile);
chomp(@noiseBarcode=<NOISEFILE>);
$expectedsize = @expectedBarcode;
$noisesize = @noiseBarcode;
close(EXPECTEDFILE);
close(NOISEFILE);
print "expected records=$expectedsize\n";
print "noise records=$noisesize\n";

print "Noise\tExpected\tDist\n";
for ($i = 0; $i < $noisesize; $i++) {

$noiseBarcode = @noiseBarcode[$i];
for ($j = 0; $j < $expectedsize; $j++) {
    $expectedBarcode = @expectedBarcode[$j];
    $dist = hd($noiseBarcode, $expectedBarcode);
    $noise = chomp($noiseBarcode);
    print "$noise	$expectedBarcode	$dist\n"
}
}

print "Expected\tNoise\tDist\n"
for ($i = 0; $i < $expectedsize; $i++) {
    $expectedBarcode = @expectedBarcode[$i];
    for ($j = 0; $j < $noisesize; $j++) {
        $noiseBarcode = @noiseBarcode[$j];
        $dist = hd($noiseBarcode, $expectedBarcode);
        $expected = chomp($expectedBarcode);
        print "$expected\t$noiseBarcode\t$dist\n"
    }
}

# ref http://www.perlmonks.org/?node_id=500235
sub hd{ length( $_[ 0 ] ) - ( ( $_[ 0 ] ^ $_[ 1 ] ) =~ tr[\0][\0] ) }
APPENDIX 2: SURVEY INSTRUMENT USED IN CHAPTER 5

Survey on risk assessment in gene therapy clinical trials

Dear Researcher,

Based on your involvement in a clinical trial of gene therapy, we would like to invite you to participate in the following 10-15 minute survey on risk assessment in clinical trials, being conducted by gene therapy researchers and ethics researchers at the Sydney Medical School, University of Sydney. We hope that this research will yield information that may be useful for other investigators involved in translational research.


This research is part of a PhD project and aims to investigate how gene therapy researchers take account of considerations of risk when making decisions about the design and conduct of gene therapy clinical trials. We wish to represent your perspective and would value your participation.

All responses will be de-identified and only de-identified information preserved, so please feel comfortable to express your views frankly. Please note that your name and email address were obtained through publicly available clinical trial registries and websites.

We would be very grateful if you could circulate the survey web address to your co-investigators and other members of your clinical trial team.

Finally, we thank you for your time and apologise for any inconvenience caused by cross-postings of this survey.

Yours Sincerely,

Associate Professor Ian Kerridge
Director, Centre for Values, Ethics and the Law in Medicine
Sydney Medical School, University of Sydney

Claire Deakin
PhD Candidate, Gene Therapy Research Unit
Sydney Medical School, University of Sydney
Risk Assessment in Clinical Trials

Introduction

This research aims to investigate how gene therapy researchers take account of considerations of risk when making decisions about the design and conduct of gene therapy clinical trials.

The questionnaire will take approximately 10-15 minutes to complete. The questions have been divided into the following sections, which will appear on separate pages.

**Part A:** demographic details (12 questions)
**Part B:** decisions about initiating a clinical trial (8 questions, 7 parts each)
**Part C:** general attitudes to research involving human subjects (5 questions)
**Part D:** attitudes to research involving human subjects based on personal experience (4 questions)
**Part E:** additional comments (optional; 1 question)

Even though you may not have been involved in trials analogous to the specific hypothetical scenarios described in Part B, we would value your judgments regarding risk-benefit assessment in these scenarios, based on your involvement and experience in research directed towards treating disease.

Thank you for giving your time to participate in our research. We greatly appreciate your thoughtful responses.
A) Demographic details

Please select the most appropriate answer.

(1) Please indicate your gender
☐ Female  ☐ Male

(2) Please indicate your age
☐ 21-30 years
☐ 31-40 years
☐ 41-50 years
☐ 51-60 years
☐ Over 60 years

(3) In which country do you conduct your research?
If your research is part of a multi-centre study, please select the country in which the centre you conduct research at is based.

☐ Australia  ☐ Czech Republic  ☐ Germany  ☐ Mexico  ☐ Russia  ☐ Switzerland
☐ Austria  ☐ Denmark  ☐ Ireland  ☐ Netherlands  ☐ Singapore  ☐ Taiwan
☐ Belgium  ☐ Egypt  ☐ Israel  ☐ New Zealand  ☐ South Korea  ☐ United Kingdom
☐ Canada  ☐ Finland  ☐ Italy  ☐ Norway  ☐ Spain  ☐ United States
☐ China  ☐ France  ☐ Japan  ☐ Poland  ☐ Sweden  ☐ Other (please specify)

(4) Have you been involved in a clinical trial in any way?
☐ Yes  ☐ No

(5) What are/were your role(s) in a current or completed clinical trial?
Select more than one description if applicable.

☐ Generated preclinical data  ☐ Vector production  ☐ Consent  ☐ Not applicable
☐ Trial design  ☐ Consent  ☐ Analysis of data  ☐ Other (please specify)
☐ Recruitment of subjects  ☐ Clinical treatment or management  ☐ Manuscript preparation
(6) What is the research setting for the preclinical research?  
Select more than one if applicable
- Hospital  
- University  
- Independent research institute  
- Other (please specify)

(7) Does the target population for your research include children?  
- Yes  
- No

(8) Are you involved in the clinical care of children, independent of any involvement you may have had in a clinical trial?  
- Yes  
- No  
- Not applicable

(9) Are you a scientist, clinician or both?  
- Scientist  
- Clinician  
- Both scientist and clinician

(10) If you answered “both clinician and scientist” in Question 9, please indicate approximately what percentage of your time is spent conducting research and what percentage of your time is spent in the clinic.  
- 10% research, 90% clinic  
- 20% research, 80% clinic  
- 30% research, 70% clinic  
- 40% research, 60% clinic  
- 50% research, 50% clinic  
- 60% research, 40% clinic  
- 70% research, 30% clinic  
- 80% research, 20% clinic  
- 90% research, 10% clinic

(11) Please select the most applicable descriptions of your field of research  
Select more than one description if applicable
- Cancer/ oncology  
- Respiratory  
- Neurology  
- Ophthalmic  
- Other (please specify)  
- Metabolic  
- Haematology  
- Neurodegenerative  
- Infectious  
- Cardiovascular  
- Immunology  
- Musculo-skeletal  
- Monogenic genetic disease

(12) How many years have you been working in this field?  
- 1-5 years  
- 5-10 years  
- 11-15 years  
- 16-20 years  
- 21-25 years  
- Over 25 years
B) Decisions about initiating a clinical trial

(1) Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of

* a statistically powered study demonstrating therapeutic efficacy in relevant primary human cells in culture AND evidence of safety in a validated cell culture assay

in the following context?

<table>
<thead>
<tr>
<th>Condition Description</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>early adulthood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is required and transplant may be associated with acute and chronic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graft-versus-host disease and other serious and secondary effects e.g. malignancy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>as acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to iron overload and cardiac failure. Life expectancy and QoL are reduced, although</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neurocognitive damage is likely. Life expectancy is normal in most patients and QoL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.
(2) Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of

a statistically powered study demonstrating therapeutic efficacy in relevant primary human cells in culture AND evidence of safety in a validated cell culture assay, if a knock-out mouse model is expected to be generated within two years

in the following context?

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.
Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of a statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no evidence of toxicity or adverse events in long term follow up, using a mouse model where the phenotype is the same as the human disease (no large animal model available) in the following context?

<table>
<thead>
<tr>
<th>Context</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.
(4) Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of a statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no evidence of toxicity or adverse events in long term follow-up, using a mouse model that involves the same gene but where the phenotype differs from the human disease (no large animal model available)

in the following context?

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.
(5) Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of

a statistically powered study demonstrating safety and therapeutic efficacy, in correcting a mouse phenotype with no adverse events, using a mouse model where the phenotype is the same as the human disease AND when a large animal model is available but a large animal study would delay trial initiation by 3 years

in the following context?

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.
(6) Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of a statistically powered study demonstrating safety and therapeutic efficacy in a large animal model, in correcting a disease phenotype that is the same as the human disease and with no adverse events in long term follow-up in the following context?

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.*
Do you think it would be appropriate to recruit subjects to a phase I trial on the basis of convincing preclinical safety and efficacy data in a mouse model AND data from a phase I trial targeting a different disease of the same target tissue, which used the same technological intervention and demonstrated therapeutic efficacy and no adverse events in the following context?

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No available treatment, presents at birth or soon after, death in infancy

No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence

No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood

No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years

Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.

Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.

Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.*
Do you think it would be appropriate to recruit subjects to a phase I trial* on the basis of convincing preclinical safety and efficacy data in a mouse model AND data from a phase I trial targeting a different disease of the same target tissue, which used the same technological intervention and demonstrated therapeutic efficacy and a low frequency of serious and life-threatening adverse events (e.g. less than 10%) in the following context?

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>No available treatment, presents at birth or soon after, death in infancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, rapidly progressive, poor quality of life (QoL), death in childhood/adolescence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, slowly progressive, poor QoL during final 10 years, death in early adulthood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No available treatment, QoL acceptable until end-stage, life expectancy reduced by 20 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be cured by allogeneic bone marrow transplantation, but prior myeloablation is required and transplant may be associated with acute and chronic graft-versus-host disease and other serious and secondary effects e.g. malignancy. Survival rate is 65% at 10 years. QoL in survivors is reduced but generally regarded as acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by regular blood transfusions, which over time may lead to iron overload and cardiac failure. Life expectancy and QoL are reduced, although QoL is considered acceptable in most cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease can be controlled by diet, but compliance is poor and low grade neurocognitive damage is likely. Life expectancy is normal in most patients and QoL is generally acceptable.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A phase I trial is the first stage of testing in human subjects and is designed to assess the safety of a new therapy.*
C) General attitudes to research involving human subjects

To what extent do you agree or disagree with the following statements?

(1) In the context of a severe disease when treatment does not exist, it is **acceptance** to proceed to clinical trials in the absence of toxicological data.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(2) The absence of adverse events in small animal models does **not** enable meaningful predictions to be made about long-term safety in humans.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(3) It is more important to have a greater prospect of benefit to participants in trials involving children than in trials involving adults.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(4) In the context of a progressive disease, it is preferable for novel therapeutic modalities with uncertain risks and benefits to be trialled initially on patients with **advanced disease** (where the patient may have 'less to lose' but also less likelihood of benefit), rather than on patients with **early disease stage** (where the patient may have a greater likelihood of benefit but also have more to lose), even when this is likely to compromise the capacity of a trial to assess efficacy.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(5) In the context of a severe disease with rapid clinical course when treatment options have been exhausted or there is no established treatment, it may be acceptable to proceed to clinical trials on the basis of a statistically powered animal study, in which endpoint measures of therapeutic benefit fall just short of clinical significance (i.e. $p \geq 0.05$).

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
D) Attitudes to research involving human subjects based on personal experience

Please answer the following questions based on your own experience.

(1) When a decision is made about commencing a trial, to what extent is the potential for adverse events to have a negative effect on public support and trust a relevant consideration?

- [ ] Completely irrelevant
- [ ] Somewhat irrelevant
- [ ] Somewhat relevant
- [ ] Very relevant
- [ ] Not applicable

(2) When a decision is made about commencing a trial, to what extent is the potential for adverse events to have a negative effect on your field of research a relevant consideration?

For example, by affecting funding, the decisions of funding bodies, or the decisions of other researchers

- [ ] Completely irrelevant
- [ ] Somewhat irrelevant
- [ ] Somewhat relevant
- [ ] Very relevant
- [ ] Not applicable

(3) Has your own personal decision-making about clinical trials ever been affected in any way by a report of an adverse event in a trial in your field?

For example, has your enthusiasm for developing, initiating or participating in a clinical trial been affected by adverse events occurring in another trial?

- [ ] Completely irrelevant
- [ ] Somewhat irrelevant
- [ ] Somewhat relevant
- [ ] Very relevant
- [ ] Not applicable

(4) Please give reasons for your response.

Optional, but any brief comments would be greatly appreciated.
E) Additional comments

Are there any additional comments or views that you would like to express?

This question has been included to give you an opportunity to express any additional comments, views or concerns that you may have about decision-making in clinical research, including comments that might arise from dissatisfaction with the present questionnaire.

It is optional, but we would greatly appreciate any comments you may have.
APPENDIX 3: PUBLICATIONS ARISING FROM THIS WORK

